Capítol 1.

Introducció.

[*Aquesta pàgina ha estat deixada en blanc intencionadament*]

La informació hereditària de tots els organismes vius, amb l'excepció d'alguns virus, està continguda en l'àcid desoxiribonucleic (ADN). Els canvis que es produeixen en la seqüència d'ADN, o mutacions, són la darrera font de variació i novetat en l'evolució dels organismes, aquests canvis poden ocórrer tant en cèl·lules somàtiques com en germinals i poden ser classificades per la longitud de seqüència d'ADN afectada. Així una mutació pot afectar un sol nucleòtid, és una mutació puntual, o varis nucleòtids adjacents, és una mutació segmental. Els canvis d'un sol nucleòtid són substitucions que es divideixen en transicions, si el canvi és entre purines o entre pirimidines, i transversions si el canvi es purines a pirimidines o al revés. Delecions i insercions són eliminacions o introduccions de nucleòtids en la seqüència respectivament. Les recombinacions ocorren quan hi ha un canvi d'un tros de seqüència per un altre i les transversions quan hi ha rotacions d'una seqüència. Aquest treball que es presenta en les següents pàgines s'ha centrat en l'estudi de les substitucions.

Poc sabem encara de les mutacions que tenen lloc en regions no codificants. Si que es poden caracteritzar les que ocorren en regions que codifiquen per proteïna segons el seu efecte en el producte final, la proteïna. Són sinònimes les mutacions que no causen canvi en l'aminoàcid codificat, altrament són no sinònimes les que si que provoquen un canvi en la seqüència final de la proteïna. En la gran majoria de vegades les mutacions sinònimes o silencioses, no alteren la seqüència final de la proteïna i no són per tant, detectables a nivell d'aminoàcid (Graur & Li, 2000). Tanmateix, en alguns casos es pot donar que una mutació sinònima no sigui silenciosa, ja que pot crear un nou lloc d'*splicing* o per contra eliminar-ne un d'existent.

Les mutacions no sinònimes que provoquen canvis aminoacídics es classifiquen en *missense* quan l'aminoàcid original és canviat per un altre. Les mutacions són *nosense* quan es canvia el codó que codifica per un aminoàcid per un codó de parada, conseqüentment la traducció finalitza prematurament i s'obté una proteïna truncada. Alternativament, mutacions en els codons stop poden provocar

que la traducció s'allargui més del compte fins i tot per sobre del lloc de poliadenilació.

Cada un dels codons que codifica per aminoàcid pot mutar a 9 altres codons, per exemple, CCU (Pro) pot experimentar sis mutacions no sinònimes, a UCU (Ser), ACU (Thr), GCU (Ala), CUU (Leu), CAU (His), o CGU (Arg) i tres mutacions sinònimes, CCC, CCA, o CCG. Com que el codi genètic universal té 61 codons amb significat, hi ha 549 possibles mutacions de substitució. En una primera aproximació, assumint que totes les mutacions són possibles en la mateixa probabilitat i que tots els codons són equiprobables en regions codificants podem derivar la probabilitat d'observar les diferents mutacions de substitució en el codi genètic (Taula 1).

Degut a l'estructura del codi genètic (consulteu apèndix per més informació) la major part de mutacions sinònimes tenen lloc en la tercera posició del codó ja que fins a un 70% d'aquestes mutacions són sinònimes. Per contra cap canvi en la segona posició és sinònima i només el 4% ho és en la primera posició.

Les mutacions de substitució es creu que tenen lloc majoritàriament en el mal aparellament de bases durant la replicació del ADN (Graur & Li, 2000). Donat que les mutacions són rares, és difícil d'estimar-ne la taxa de manera directa i s'ha de fer de manera indirecta. Si que se coneix la taxa de mutació del ADN nuclear en mamífers i varia molt segons les zones genòmiques que estudiem (Graur & Li, 2000). Tot i així, aquesta taxa és unes 10 vegades més baixa que la del ADN mitocondrial. La taxa de mutació d'organismes amb ARN, bàsicament virus, és varis ordres de magnitud major degut a que les polimerases d'aquests organismes són molt propenses a cometre errors i els sistemes de correcció i reparació no són tant eficients (Graur & Li, 2000).

La direcció de les mutacions no és aleatòria, de fet les transicions són entre un 60 i un 70% de totes les mutacions observades en genomes nuclears de cèl·lules animals, mentre que la proporció que hauríem d'observar sota condicions

aleatòries seria només del 33% (4 transicions sobre 12 transversions observables).

Taula 1 (Graur & Li, 2000)

Substitucions	Total	Percentatge
Total en tots els codons	549	100
Sinònims	134	25
No Sinònims	415	75
Missense	392	71
Nonsense	23	4
Total en primera posició	183	100
Sinònims	8	4
No Sinònims	175	96
Missense	166	91
Nonsense	9	5
Total en segona posició	183	100
Sinònims	0	0
No Sinònims	183	100
Missense	176	96
Nonsense	7	4
Total en tercera posició	183	100
Sinònims	126	69
No Sinònims	57	31
Missense	50	27
Nonsense	7	4

En genomes mitocondrials aquestes diferències esdevenen encara més grans. De fet, alguns nucleòtids són més mutables que altres i en ADN de mamífers es troba que G i C muten de manera més freqüent que A o T.

Tampoc les mutacions són aleatòries pel que fa a la posició en el genoma, hi ha posicions que muten més que les altres són el que anomenem *hot spots* o punts calents (Graur & Li, 2000).

I. EL CONTEXT EVOLUTIU DE LES MUTACIONS

És crucial entendre el paper que juguen les mutacions en l'evolució, l'adaptació de les espècies i les raons per les quals mutacions nocives o potencialment perjudicials són presents en els diferents organismes. És important entendre en aquests casos el context evolutiu en el que apareixen i ens ho permet en part el marc contextual que ofereixen la genètica de poblacions i les teories evolutives.

Les mutacions poden o no afectar el genotip i d'acord amb això la selecció actuarà sobre elles. Si el genotip es veu alterat es pot afectar la salut de l'individu, són mutacions patològiques (en anglès *deleterious*), contra elles actuarà la selecció i seran finalment eliminades de la població. Aquesta selecció s'anomena negativa o *purifying selection*. Ocasionalment una mutació pot ser tant bona com el millor al·lel de la població, és una mutació selectivament neutra i el seu destí no estarà regit per la selecció. Finalment en molt pocs casos, una mutació pot arribar a millorar la salut o *fitness* de la població, és una mutació avantatjosa i serà sotmesa a una selecció positiva.

Malgrat tot la selecció natural no és l'únic factor que pot causar canvis en les freqüències al·lèliques. Aquestes també poden canviar per atzar, i ara els canvis no són dirigits sinó que són aleatòris, és la deriva genètica aleatòria (*random genetic drift*). Un cas clar de deriva genètica bé donada pel desequilibri existent durant la reproducció entre el número de mostres possibles (gàmetes) i el número d'individus que és molt menor. L'atzar guia la producció de gàmetes. La usual diferència en el nombre de gàmetes masculins i femenins en la fecundació, dóna un joc de probabilitats, que mostren la importància de la deriva genètica en el balanç final al llarg de les generacions. D'aquesta manera les freqüències dels al·lels s'alteren sense intervenció de la selecció (Graur & Li, 2000).

Les freqüències dels al·lels varien de generació en generació fins que arriba a 0, que és la pèrdua o extinció d'aquest, o bé arriba a 1 aleshores parlem de fixació.

Aquest és el destí final de qualsevol al·lel, la deriva genètica porta a l'extinció d'uns i la fixació d'altres sempre que hi hagi prou temps. Ara bé, la dinàmica de fixació o extinció d'al·lels varia molt segons siguin mutacions avantatjoses o neutres, mentre que les primeres s'eliminen o es fixen molt ràpidament, les segones tenen un comportament més erràtic i el temps que necessiten per fixar-se o desaparèixer és molt més llarg.

La importància que es dóna a la selecció i a la deriva genètica fa que s'expliqui de manera diferent el fet de l'evolució. Així la teoria mutacionista explica l'evolució per l'entrada de mutacions i la deriva genètica. La teoria seleccionista dóna més pes a la selecció positiva i balancejada. La teoria neutralista desenvolupada per Kimura (Kimura & Takahata, 1983) dóna més pes a la mutació, deriva genètica i la selecció per purificació (Graur & Li, 2000). Per aquesta última els polimorfismes són inestables i temporals i estan en procés d'extinció o fixació. Tanmateix existeix selecció positiva i direccional en alguns gens (Fay et al., 2001), així com al·lels lleugerament patològics i selecció balancejada (Graur & Li, 2000).

Darrerament, degut a diferents projectes d'anàlisi de SNPs (*Single Nucleotide Polimorphism,* polimorfismes d'un sol nucleòtid; són varicions en les seqüències d'un sol nucleòtid) (Cargill et al., 1999; Collins et al., 1998; Halushka, 1999; Lander et al., 2001; Sachidanandam et al., 2001; Sunyaev et al., 2000c; Syvanen, 2001; Venter et al., 2001) s'està fent una revisió de les diferents teories i del paper de la selecció positiva, selecció negativa, la deriva genètica, els haplotips de SNPs (Salisbury et al., 2003) i el paper de les malalties (Fay & Wu, 2001; Fay & Wu, 2003; Fay et al., 2001; Hacia et al., 1999; Lercher & Hurst, 2002; Sunyaev et al., 2003; Wong et al., 2003).

Tot això mostra que en general la teoria neutralista es compleix i la majoria de mutacions apareixen per raons d'atzar i les fixen efectes probabilístics però també existeixen mutacions que no es regeixen per aquestes variables i que tenen un efecte sobre l'individu. Aquestes mutacions explicarien patologies i

sensibilitats especials a problemes de salut. Algunes sembla ser que estarien en procés de desaparició però d'altres es conserven en els genomes per raons que tot just ara es comencen a conèixer. Caldria doncs estimar la naturalesa de les mutacions i les forces que les guien.

II. MUTACIONS PUNTUALS/SNPs

La culminació en els darrers anys dels diferents projectes genoma i especialment del genoma humà han aportat a la biologia i a la medicina una quantitat ingent de dades que tot just estan en procés d'anàlisi. Entre aquestes dades cal destacar els SNPs, que són les diferències d'una sola base que es donen al comparar cromosomes d'una mateixa espècie. Els SNPs són els responsables del 90% de la diversitat fenotípica entre individus d'una espècie. Aquesta variabilitat explicaria també el risc a sofrir malalties i la resposta a factors ambientals (Cargill et al., 1999). Els primers resultats derivats del projecte genoma huma mostren que l'any 2001 ja hi ha catalogats més 1.42 milions de SNPs amb una freqüència de 1 SNP cada mil bases (Sachidanandam et al., 2001). L'interès pels SNPs és degut en gran part a que es creu que la gran majoria de variabilitat és deguda a variants freqüents, això són variants amb una freqüència superior a l'1% en la població general. Les variants observades en freqüències inferiors al 1% es solen anomenar variants rares i no es solen considerar SNPs a efectes pràctics.

Els primers estudis fets sobre SNPs en poblacions grans i en gens d'interès biomèdic mostren que variants molt rares explicarien les malalties hereditàries menys comuns causades per un sol gen i que s'hereden segons un patró mendelià clàssic. Les variants més comunes d'aquests polimorfismes explicarien la gran majoria d'heterozigositat present en les poblacions. Així la combinació de diferents polimorfismes comuns donarien la susceptibilitat a malalties complexes que no s'hereden segons el model mendelià simple i que afecten a un percentatge

elevat de la població. Per aquest motiu és important associar aquesta variació en la heterozigositat amb la variació fenotípica i d'aquesta manera detectar gens de susceptibilitat a aquestes malalties.

La medicina actual ha posat moltes esperances en comprendre millor els SNPs i com es distribueixen en la població. Això ens pot permetre d'identificar i conèixer millor gens responsables de susceptibilitats a malalties. Un model d'estudi idoni és el de les mutacions puntuals en proteïnes responsables de malalties hereditàries monogèniques. Aquestes malalties són molt poc comuns i es solen presentar en diferents individus dins una mateixa família. Per aquesta raó i que s'hereden segons un patró mendelià es va observar que eren causades per mutacions en un sol gen. En molts casos estan causades per mutacions puntuals molt poc freqüents (alguns autors xifren la seva freqüència sempre per sota del 1%). L'estudi d'aquestes malalties per part de la comunitat científica ha estat intens i ha permès, usant tècniques de genètica evolutiva i mapatge genètic, identificar els gens "responsables". L'anàlisi bioinformàtica de totes aquestes mutacions ens pot fer entendre millor quins canvis són patològics i quins neutres i racionalitzar les raons físico-químiques, estructurals i evolutives que provoquen aquestes diferències (Ferrer-Costa et al., 2002; Ng & Henikoff, 2001; Saunders & Baker, 2002; Sunyaev et al., 2000a; Wang & Moult, 2001) .

Tal com s'ha dit anteriorment amb l'aparició d'esborranys de genomes complets com els de l'home (Lander et al., 2001; Venter et al., 2001) o ratolí (Waterston et al., 2002) així com d'altres espècies ha permès tenir per primera vegada dades en quantitat per poder estudiar a fons les diferents característiques de les mutacions puntuals.

Diferents treballs es fan ressò de la variabilitat intraespecífica usant diferents aproximacions com, seqüenciar diferents grups de gens d'interès biomèdic en diferents grups poblacionals (Cargill et al., 1999; Halushka, 1999), o bé analitzant la variabilitat en bases de dades de ESTs humanes (Sunyaev et al., 2000c) o analitzant bases de dades de SNPs (Zhao et al., 2003). Aquests autors

mostren que en general els SNPs es troben en la mateixa freqüència tant en zones codificants com en zones no codificants i al voltant de 1 cada 350 parell de bases (Cargill et al., 1999). Dins els gens, aproximadament la meitat dels SNPs són no sinònims, fet que mostra una forta selecció en contra dels canvis no-sinònims i més si tenim en compte que aquests solen aparèixer en les freqüència poblacionals més baixes.

Aquesta selecció en contra de certes variants fa plantejar que, malgrat que algunes variants seran neutres per la salut, d'altres seran patològiques o almenys provocaran alguna alteració del funcionament normal de la proteïna. Comprendre les desavantatges funcionals o estructurals que fan que moltes mutacions no sinònimes siguin eliminades o estiguin en procés d'eliminació ha de permetre detectar i esmenar alteracions que poden afectar a la salut de l'individu.

III. MUTACIONS I PATOLOGIA

Sembla clar avui que una part de les nostres malalties és ocasionada per mutacions puntuals que afecten a la seqüència de la proteïna. Entre aquestes es diferencien les malalties monogèniques i les poligèniques. Les malalties monogèniques o mendelianes tenen un patró hereditari definit per les lleis de la genètica clàssica. Els estudis de desequilibri de lligament mostren que aquestes malalties són produïdes per mutacions en un sol gen. En certs casos aquestes mutacions són puntuals i s'ha esdevingut el canvi d'un aminoàcid per un altre resultant alterada de manera nociva la funció de la proteïna. En general aquestes mutacions tenen una freqüència molt baixa en població, per regla general per sota del 1%. En alguns casos, però aquesta freqüència està enriquida en certes poblacions, és el cas, per exemple, de les mutacions que afecten a l'hemoglobina i que donen lloc a l'anèmia falciforme o alguns tipus de talassèmies. Aquestes variants presenten freqüències elevades en poblacions africanes i s'ha relacionat

amb la resistència a l'infecció per *Plasmodium falciparum* agent causant de la malària (Zuckerkandl & Pauling, 1962). Aquest factor extern o ambiental fa que sigui una avantatge ser portador d'aquesta mutació patològica, almenys en heterozigositat.

Les malalties més complexes que no s'hereden segons un patró clàssic mendelià són més difícils d'estudiar i de relacionar-les amb un sol gen. En general es relacionen amb diversos gens. El més acceptat fins ara, és que l'acumulació de mutacions poc greus en diferents gens produeix l'aparició d'aquestes malalties que són molt esteses en la població, en són exemples, la diabetis, la hipertensió o les malalties coronaries. Alguns autors defensen la teoria de la variant comuna malaltia comuna (en angles CV-CD *common variant-common disease*) (Collins et al., 1997; Lander, 1996; Risch & Merikangas, 1996), segons la qual alguns SNPs que són presents en població en freqüències més elevades marcarien el risc genètic a sofrir malalties. Així l'esperança de poder descobrir nous gens relacionats amb malalties comuns passa per fer estudis d'associació de SNPs en el genoma, d'aquí que hi hagi una raó més per generar mapes d'alta densitat de SNPs amb estudis poblacionals.

En els dos tipus de malalties hi ha un interès per saber quines mutacions o SNPs són neutres i quins altres són patològics o lleugerament patològics. Tal i com hem vist fins ara, extreure informació funcional d'un SNP no esdevé trivial des del punt de vista de l'anàlisi genòmica, cal doncs buscar nous camins, que han de passar per força per comprendre les característiques estructurals, funcionals i energètiques de les proteïnes i el seu entorn cel·lular.

IV. BASES MOLECULARS DE L'IMPACTE FUNCIONAL/ ESTRUCTURAL DE LES MUTACIONS

En els anys 60 Zuckerkandl i Pauling parlaven de tres tipus de malalties moleculars (Zuckerkandl & Pauling, 1962); les mutacions en una proteïna podien afectar a la funció molecular, a la interacció amb l'entorn intracel·lular o bé interferir en el nivell de síntesi.

Entre les primeres, hi ha aquelles mutacions que alteren centres funcionals de les proteïnes, ja sigui el centre actiu o be altres centres secundaris relacionats amb la regulació o bé reconeixement. Entre les mutacions que afecten la interacció amb l'entorn cel·lular hi ha aquelles que afecten especialment la solubilitat. Atenent a que les proteïnes tenen en termes generals un gran nombre de residus apolars, aconseguir durant el plegament una estructura soluble no esdevé una tasca fàcil. Així mutacions que alterin aquest equilibri precari, donen lloc a estructures poc solubles (Zuckerkandl & Pauling, 1962). Altres alteracions a la superfície provoquen formació d'estructures supramoleculars com fibriles o plaques.

Finalment la interferència en la síntesi també poden provocar malalties monogèniques. Les causes que ho provoquen poden ser alteracions en les regions reguladores del gen, alteracions en la síntesi i processat dels missatgers, de la proteïna, alteracions en el plegament o en el procés de control de qualitat o alteracions en el procés de transport ja sigui intra com interce·lular. També provoquen malalties d'aquest tipus l'alteració dels nivells de síntesi de les diferents subunitats que conformen estructures quaternàries que comporta una producció global alterada.

Aquesta classificació ens permet racionalitzar i encarar l'estudi l'efecte final de les mutacions puntuals sobre la proteïna i sobre l'entorn proteic, per tal de poder entendre quan algunes mutacions provocaran malalties.

El primer nivell de comprensió és sens dubte l'efecte de la mutació sobre la proteïna des d'un punt de vista estructural com energètic. Cal tenir en compte també, l'efecte de l'entorn cel·lular. Les proteïnes desenvolupen la seva tasca en un entorn altament poblat on la mobilitat i l'accés a partícules de solvent està altament limitada i els problemes d'agregació per interacció amb altres macromolècules és constant. Aquest entorn s'assembla poc al d'assaigs *in vitro* sobre funció i estructura que es realitzen en solucions molt diluïdes de proteïna.

L'efecte estructural de les mutacions puntuals

Durant els anys 80 i principis dels 90 del segle XX va ser intensa l'activitat de recerca en el camp de la cristal·lografia de proteïnes en l'anàlisi de mutants puntuals i en la deriva a partir d'aquests de paràmetres energètics i estructurals. Els coneixements previs de plegament de proteïnes i estabilitat proteica combinat amb l'estudi d'aquests mutants cristal·litzats així com la comparació estructural entre ells va permetre establir les bases del coneixement de l'efecte de les mutacions puntuals sobre l'estructura de la proteïna.

Sembla clar que les proteïnes són marginalment estables, hi ha un balanç molt delicat entre les interaccions estabilitzants i les desestabilitzants i amb una diferència energètica entre les formes plegades o desplegades entre les 5-20 kcal/mol. Per tant, petits canvis en les contribucions d'estabilització com de desestabilització poden alterar substancialment el resultat final(Matthews, 1987).

Diferents treballs (Fersht & Serrano, 1993; Matthews, 1975; Matthews, 1987; Matthews, 1993; Matthews, 1995; Poteete et al., 1992; Poteete et al., 1997; Rennell et al., 1991; Shortle, 1992) sobre estabilitat de diferents mutants en una mateixa proteïna han permès establir tendències en els efectes de les diferents mutacions tant en l'estabilitat com en la funció de la proteïna. Així més concretament, els treballs de Matthews (Matthews, 1995) i Poteete (Rennell et

al., 1991) que fan mutació exhaustiva de lisozim de bacteriòfag T4 mostren que en general la majoria de substitucions tenen un efecte petit en l'estabilitat de les proteïnes i moltes conserven certa activitat comparable amb la proteïna salvatge. Les conseqüències d'una mutació depèn de la naturalesa del canvi i de l'entorn on la mutació té lloc.

A nivell estructural s'ha observat que les mutacions estabilitzants i desestabilitzants creen alteracions quasi inapreciables en l'estructura global de la proteïna. Localment les alteracions són petites, generalment els àtoms de les cadenes laterals en la zona de la mutació poden sofrir canvis de fins a 1 Å, i els reordenaments de la cadena principal no són superiors de 0.5Å (Matthews, 1995). En general no s'observen canvis més enllà de 5-10 residus de distància ni a més de 10 Å al voltant del punt de mutació. També s'ha vist que les mutacions més desestabilitzants solen a ocórrer en zones de baixa mobilitat i baixa accessibilitat al solvent. En el cas de lisozim (Matthews, 1995), els autors mostren que les mutacions en superfície tendeixen a ser poc lesives i en general canvis per alanina quasi no tenen cap efecte en la proteïna. Tanmateix cal assenyalar que algunes mutacions es relacionen amb l'estat desplegat. Es creu que augmenten l'estabilitat de l'estat desplegat i aquestes poden trobar-se en qualsevol lloc, fins hi tot en les zones més mòbils.

En alguns casos l'efecte de les mutacions pot ser bastant subtil, per exemple, se sap que algunes proteïnes tenen diferents confòrmers de baixa energia que representen l'estructura salvatge, alguns mutants poden afavorir o estabilitzar un confòrmer respecte l'altre (Matthews, 1993). En particular, en alguns casos la diferència entre confòrmers és també funcional, sent un d'ells funcional i l'altre no, en aquests casos aquest tipus de mutants tindran sens dubte efecte en la funció final de la proteïna, sense que hi hagi un efecte directe sobre l'estructura. Aquests efectes s'accentuen en sistemes més complexos com l'al·losterisme o bé estructures quaternàries i en interacció entre proteïnes i proteïna/lligand. En tots aquests casos mutacions que afecten molt poc o gens l'estructura de la proteïna

poden tenir efectes dramàtics en el funcionament de final del sistema. Molts d'aquests mutants són mutants en superfície que alteren la forma o bé les propietats físico-químiques de la superfície de reconeixement.

Hi ha mutacions que afavoreixen la formació d'agregats durant els processos de plegament i desplegament de les proteïnes. En aquests casos sembla que el paper de la mutació no és tant la desestabilització com el d'afavorir la formació d'agregats en l'estat desplegat (Ramírez-Alvarado & Regan, 2002).

Un aspecte important de les mutacions puntuals és el seu efecte sobre els aspectes dinàmics de l'estructura de la proteïna L'aplicació de metodologies de la química computacional, especialment de la dinàmica molecular es poden simular els comportaments dinàmics de diferents proteïnes. Ens els recents anys i gràcies a l'augment de la potència de càlcul dels ordinadors i al desenvolupament d'algoritmes de paral·lelització es poden allargar els temps de simulació. Alguns treballs aborden ja la simulació amb tècniques de dinàmica molecular el problema de les mutacions puntuals relacionades amb patologia (Cox et al., 2003; El-Bastawissy et al., 2001; Futatsugi & Tsuda, 2001; Perryman et al., 2004; Stockner et al., 2003). En tots aquests treballs es mostren que hi ha certs comportaments diferencials entre la proteïna salvatge i la mutant que podrien explicar el comportament patològic de la proteïna.

L'efecte de l'entorn cel·lular sobre l'efecte de les mutacions puntuals

Fins aquí hem vist com afecten les mutacions a l'estructura i a la dinàmica de les proteïnes, tots aquests experiments però s'han fet en condicions in vitro. Tanmateix no es pot negligir que la vida té lloc a dins la cèl·lula i aquest entorn és pel que sabem avui bastant diferent del simulat en les condicions in vitro. Estudis començats a finals dels anys 80 es comencen a interessar per les

condicions d'entorn que afecten l'interior de la cèl·lula. Els autors entre ells Minton (Minton, 2000) i Ellis (Ellis, 2001; Ellis & Minton, 2003) introdueixen el concepte de *molecular crowding* que s'aplica a sistemes biològics i descriu el fet que la concentració total de macromolècules és tan elevada dins la cèl·lula que una proporció significant del volum és físicament ocupada i no és accessible a altres molècules. Així, la concentració efectiva i per tant l'activitat termodinàmica de cada espècie macromolecular dins la cèl·lula és major que la concentració real i aquesta diferència té conseqüències en les propietats cinètiques i termodinàmiques. S'estima que la concentració real de proteïna és de 200-300 g/l i la d'àcids nucleics entre 70-150 g/l. La sang sembla que conté fins a 350g/l d'hemoglobina. A l'entorn extracel·lular, els glúcids de la matriu també contribueixen a aquest mateix efecte. En termes generals entre un 20-30% del volum cel·lular és ocupat per macromolècules.

La teoria que recolza el *molecular crowding* prediu que augmenta dos aspectes del procés del plegament de proteïnes. Per una part s'accelera el primer col·lapse de la cadena polipetídica ja sigui en el plegament de la nova cadena sintetitzada o bé desplegada. Però també s'accelera la formació d'agregats no funcionals de proteïnes en procés de plegament. Les proteïnes mig plegades tendeixen a agregar-se entre elles i formen agregats no funcionals que solen precipitar (Ellis, 2001).

En relació amb les patologies, cal assenyalar l'efecte que pot tenir el *crowding*, així la formació de fibres d'anèmia falciforme s'accelera en presència d'elevades concentracions de la hemoglobina S desoxigenada(Minton, 2000).

Sembla que la formació de fibriles amiloides en malalties neurodegeneratives com la malaltia d'Alzheimer podria estar afavorida per fenòmens de *crowding* com els esmentats.

Per tant, no es pot fer cap afirmació categòrica sobre l'efecte de l'entorn sobre una mutació puntual en una proteïna, però si que podria ser que canvis que en un

entorn *in vitro* aparentment no tenen cap efecte, si que el tinguin en un entorn *in vivo*.

V. MUTACIONS PATOLÒGIQUES

Tal com s'ha comentat anteriorment, les mutacions poden causar dos tipus de patologies, monogèniques o poligèniques. El cas de les malalties poligèniques o complexes és poc conegut, ja que no estan identificats la totalitat de gens implicats ni totes les mutacions que afecten els pacients ni l'impacte dels efectes ambientals en el desenvolupament de les malalties. El cas de les malalties monogèniques és més simple, ja que en molts casos es coneix el gen implicat i també les mutacions causants de les malalties. L'estudi de les raons físiques i químiques, estructurals i cel·lulars que porten una mutació d'aquest tipus a esdevenir patològiques constitueix un primer i valuós pas en la comprensió de la base molecular de la malaltia.

Tal i com s'ha mencionat anteriorment les mutacions afecten al correcte funcionament de les proteïnes a tres nivells, a nivell d'afectació de les regions relacionades amb la funció pròpiament dita, a la seva interacció amb l'entorn i finalment a la seva síntesi. A continuació es mostren diferents casos de mutacions puntuals causants de malalties monogèniques que intenten mostrar aquestes idees.

-ANÈMIA FALCIFORME

El cas de l'anèmia falciforme és paradigmàtic entre les malalties monogèniques per ser el primer que es va caracteritzar i també per desenvolupar-se durant el seu estudi un seguit de tècniques bioquímiques que van marcar l'evolució de la biologia molecular dels anys 80 i 90.

La malaltia es va descriure per primera vegada el 1904 a Chicago per Irons i Herrick (Herrick, 1910; Savitt & Goldberg, 1989), que estudiant la citologia d'una mostra de sang d'un pacient es va adonar de la presència d'una notable quantitat de cèl·lules amb "forma allargada i de pera". El pacient que va sofrir de forma recurrent en episodis de malaltia durant els següents 11 anys va morir a l'edat de 32 anys. Posteriorment, es va saber que la malaltia era causada per una forma anòmala de hemoglobina, que és la proteïna responsable del transport de l'oxigen en la sang dins dels eritròcits.

Es va mostrar que la malatia era recessiva i que l'homozigot presentava una anèmia total que era letal a principis del segle XX i que actualment tot i els tractaments és una malaltia greu. L'heterozigot, no pateix greus alteracions de la salut excepte en condicions altament estressants. Es va descobrir que la malaltia es presentava de manera heterozigota en una freqüència molt elevada entre la població negra a Àfrica i a Amèrica.

Va ser el 1952 quan Vernon Ingram (Ingram, 2004) va començar a treballar en la seqüenciació de la proteïna de l'anèmia falciforme ja que poc temps abans Perutz havia mostrat la baixa solubilitat de l'hemoglobina desoxigenada de pacients d'anèmia falciforme. Adicionalment, Pauling, el 1949 (Pauling L, 1949) havia demostrat que les hemoglobines salvatge i mutant tenien mobilitat electroforètica diferent i com que aquesta propietat depenia de l'estructura proteica, aquesta havia de ser diferent. Ingram va aconseguir demostrar de manera indirecta que la única diferència entre la proteïna salvatge i la malalta era un sol aminoàcid. Posteriorment es va demostrar amb la publicació de les dues seqüències completes que les suposicions de Ingram eren les correctes (Ingram, 2004; McKusick-Nathans Institute for Genetic Medicine, 2000). El primer impacte conceptual va ser que un sol canvi aminoàcidic provoqués un canvi de comportament tant gran en la proteïna i conseqüentment a la cèl·lula i a l'individu. Actualment tenim una imatge molt més clara del problema. La mutació de l'anèmia falciforme té lloc al gen de la cadena β de l'hemoglobina.

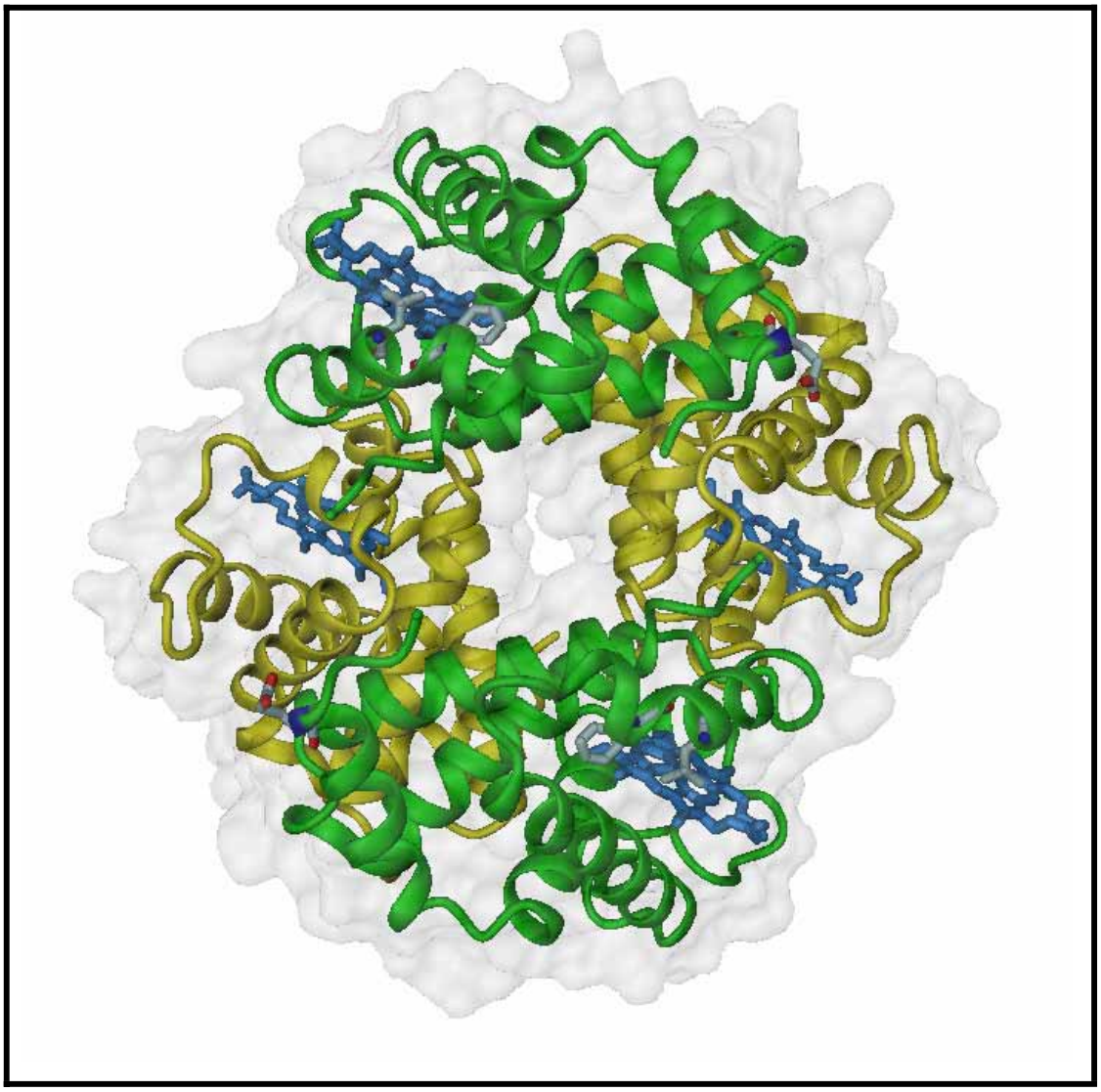

Fig1. Hemoglobina (codi pdb 1A3N). S'observa l'estructura quaternaria $\alpha_2\beta_2$, les cadenes α en groc, les cadenes β en verd. En blau intens el grup hemo. Dins la cadena β en blau clar el residu de glutàmic a la posició 6 situat prop de la cadena α i els residus 85 i 88, que formen la cavitat hidrofòbica prop del grup hemo.

L'estructura quaternaria funcional de l'hemoglobina en un individu adult és un tetràmer format per dues cadenes alfa i dues cadenes beta ($\alpha_2\beta_2$), tal i com es pot veure en la figura1. Les cadenes alfa estan codificades per dos gens diferents els alfa1 i alfa2 mentre que les beta codifiquen en el gen beta, aquests gens es troben en un *cluster* on hi ha altres gens i pseudogens que deriven de diferents processos de duplicació gènica. L'hemoglobina estudiada per Ingram és coneguda per

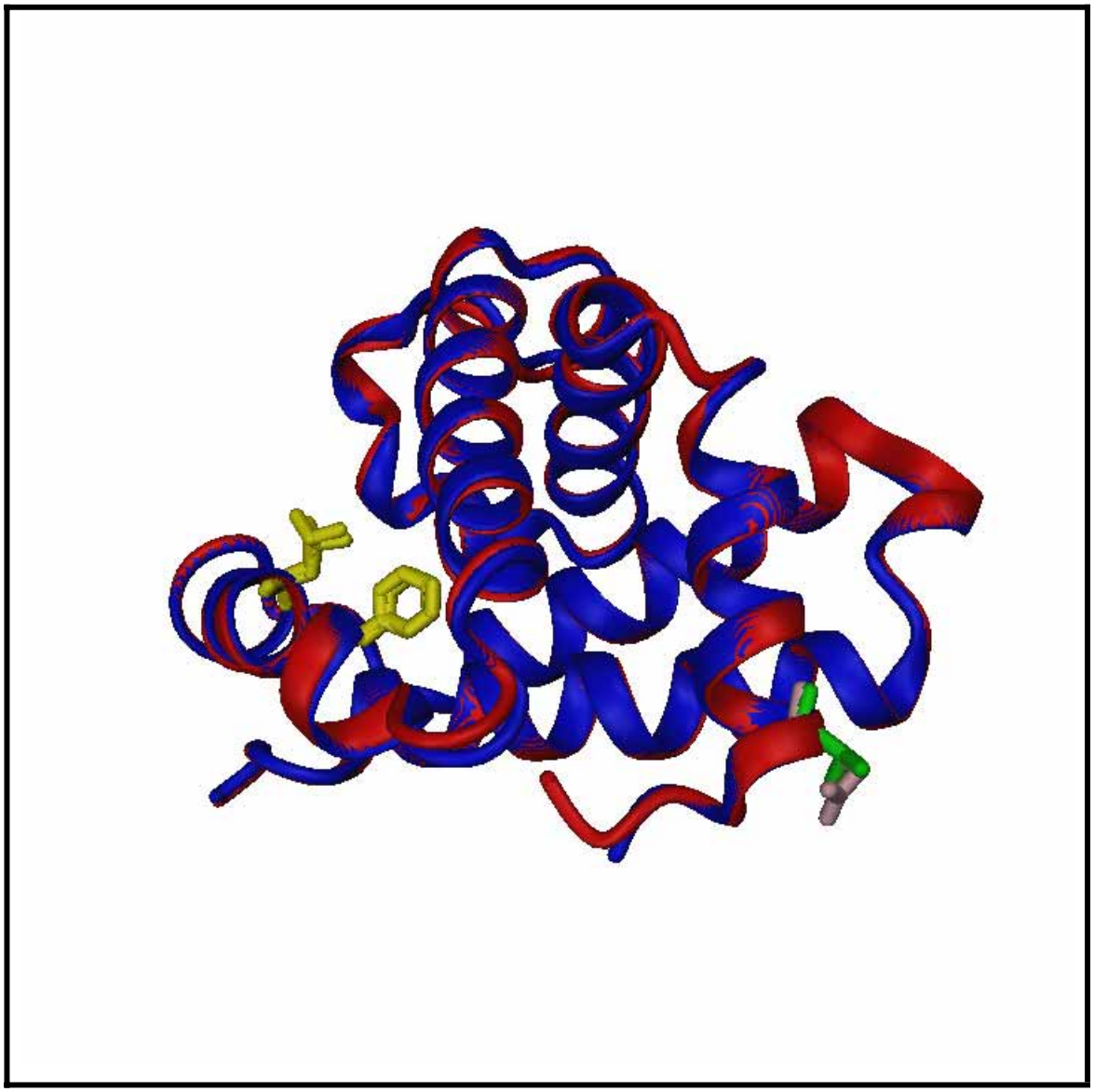

Fig2.Superposició de les cadenes β salvatge en color blau (codi pdb 1A3N) i mutant en color vermell (codi pdb 2HBS). En gris residu glutàmic de la posició 6 que muta a valina en verd. En groc residus phe 85 i leu 88 que formen el *pocket* hidrofòbic on interaccionarà la valina mutant en la forma desoxigenada de la hemoglobina.

hemoglobina S i conté una mutació en la posició 6 de la cadena beta, provocant el canvi de glutàmic a valina. La superposició de l'estructura salvatge i mutant mostren que pràcticament no s'altera gens l'estructura tridimensional de la proteïna (veure figura 2), bàsicament degut a que la posició 6 es troba a l superfície i prou distanciada de la zona d'interacció amb el grup hemo i de la

zona d'interacció amb les altres hemoglobines en l'estructura quaternària (veure figura 2 i figura 1).

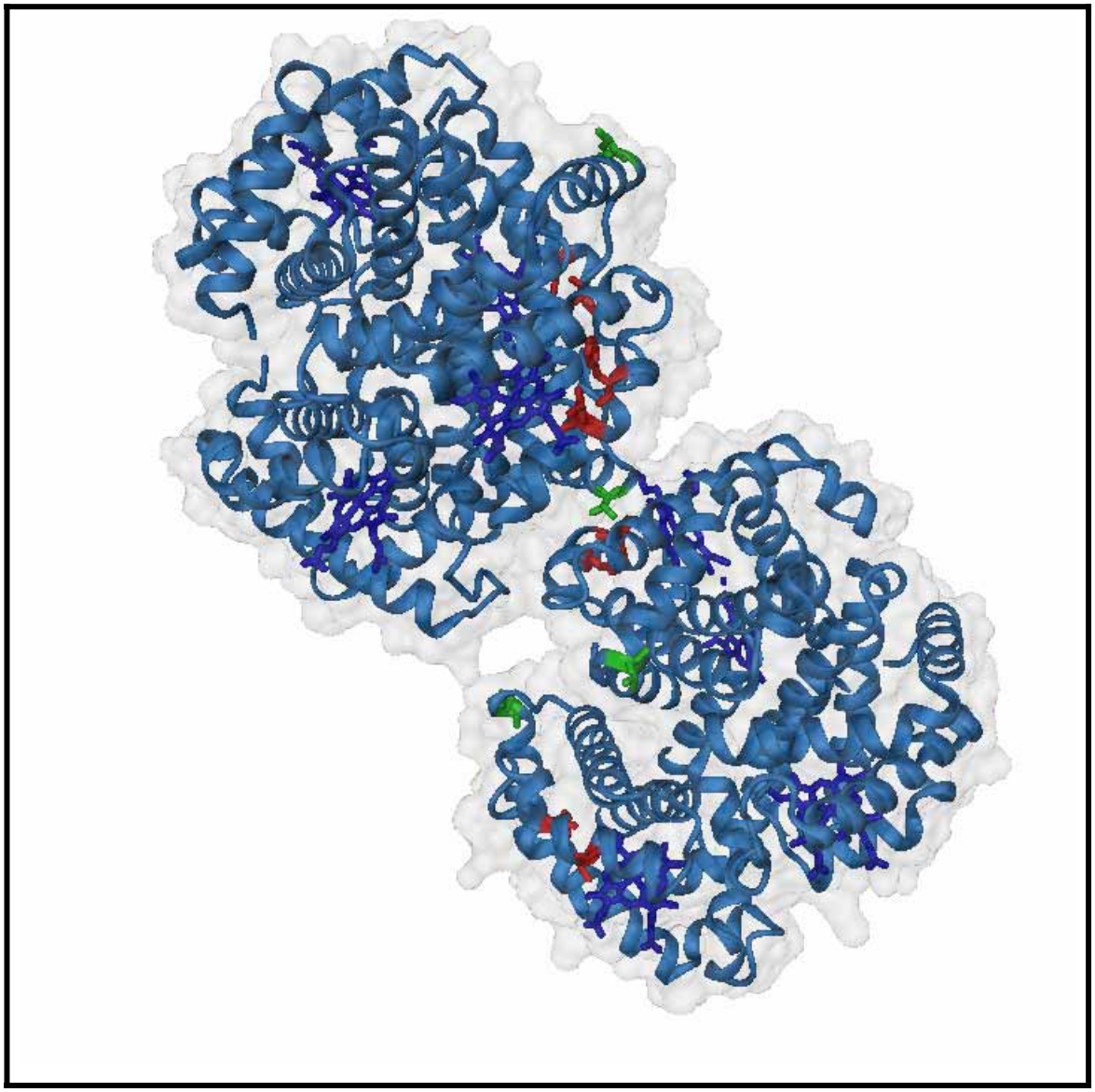

Fig3. Visió general de dues hemoglobines S desoxigenades formant interacció entre la val 6 d'una de les cadenes β d'una hemoglobina amb el *pocket* hidrofòbic format per phe 85 i leu 88 d'una de les cadenes β de la hemoglobina següent.

Els estudis de cristal·lografia mostren que principalment la cadena lateral d'aquesta nova valina pot inserir-se dins una cavitat hidrofòbica formada per la leucina 88 i la fenilalanina 85 d'una hemoglobina adjacent que estigui desoxigenada. La cavitat hidrofòbica es pot apreciar en la figura 2. La interacció entre hemoglobines mutants desoxigenades es pot observar en la figura 3.

Aquesta interacció genera la formació de fibres allargades de proteïna dins dels eritròcits que li confereixen aquesta forma de falç allargada característica (Huisman et al., 1996) (veure figura 3,4 i 5).

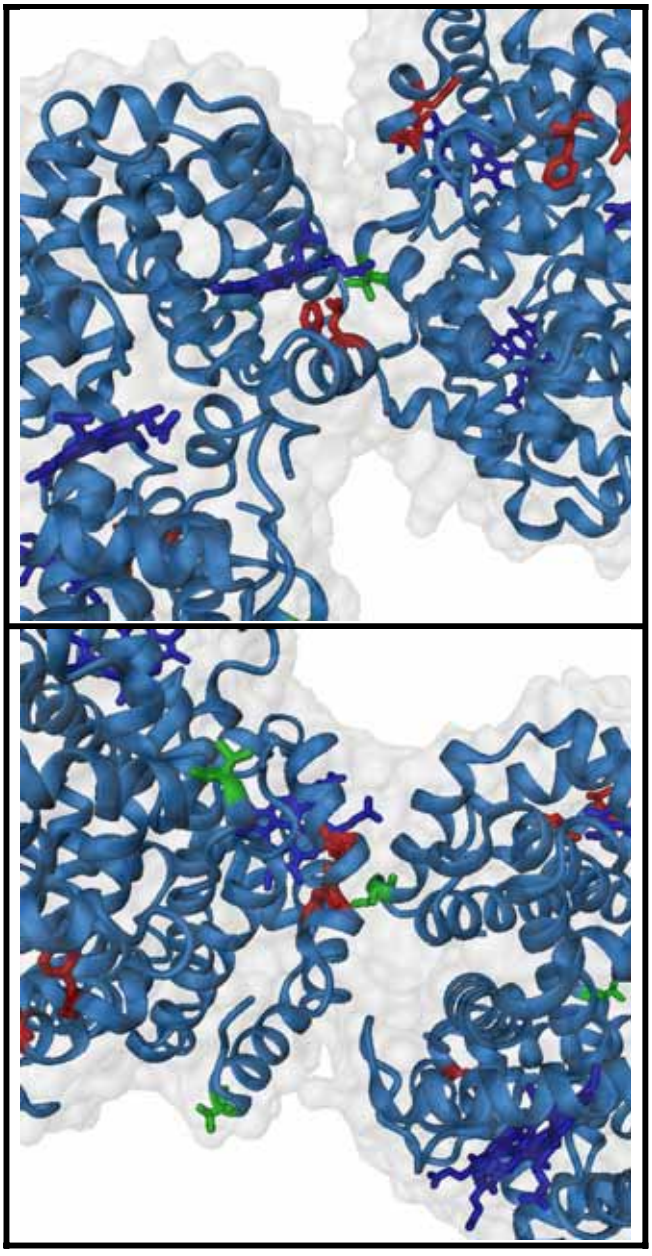

Fig4 i 5. Imatge detallada de la interacció val6 de la cadena β amb la cavitat hidrofòbica phe85 leu 88 de la cadena β de la següent hemoglobina.

Tal i com reconeix Ingram (Ingram, 2004), només l'atzar va fer que justament es donés el tipus de canvi en el lloc adequat i que això provoqués l'agregació i

formació de fibres. Si això no hagués estat així segurament hauria estat un canvi neutre sense efecte ni història en la biologia.

Actualment hi ha diferents bases de dades que recullen noves mutacions en hemoglobines que són patològiques en diferents graus. De mutacions puntuals que causen anèmia falciforme se'n comptabilitzen fins a 630 i d'elles 530 corresponen a les cadenes α i β les més abundants en individus adults (Huisman et al., 1996). D'aquestes se'n desprèn la certesa que mutacions que en principi es creien silencioses alteren l'estabilitat o les propietats físico-químiques de la proteïna. Així aquestes mutacions impedeixen totalment la seva detecció per mètodes clàssics a nivell de proteïna ja sigui perquè ni tan sols arriba a existir la proteïna funcional o perquè les seves propietats electroforètiques són exactament iguals que les formes salvatge (Huisman et al., 1996).

Analitzant la distribució de mutacions en les cadenes alfa i beta ens adonem que no hi ha segments més sensibles a mutació que els altres. Les anomalies que es detecten són canvis en la funcionalitat o les propietats físico-químiques que porten a formació de fibres, alta afinitat per l'oxigen, eritrocitòsis i inestabilitat.

El cas de l'hemoglobina S estudiat per Ingram és un cas clar de l'alteració de la interacció de les proteïnes amb el seu entorn. Així una mutació en superfície genera un nou centre d'interacció amb un residu d'una proteïna veïna que ocasiona un canvi dràstic en la solubilitat de la proteïna desoxigenada per formació de fibríles.

-AGAMMAGLOBULINÈMIA LLIGADA AL X

Els exemples estudiats en aquesta malaltia s'emmarquen dins el grup de mutacions que afecten directament la funcionalitat de la proteïna ja que s'afecten zones d'interacció amb lligands, amb altres proteïnes o centres actius.

L'agammaglobulinèmia és una immunodeficiència lligada al cromosoma X caracteritzada per una manca de maduració de linfòcits B associada amb una manca de rearranjaments de la cadena pesant de immunoglobulines Ig. El defecte s'ha associat a la Bruton tirosina cinasa, que és un regulador clau en la maduració dels linfòcits B.

Els pacients d'aquestes malalties són molt sensibles a infeccions bacterianes però no de virals. En molts pacients apareixen quadres d'artritis reumatoides. En la forma més usual de la malaltia, no hi ha limfòcits B en plasma. Abans de l'aparició dels antibiòtics la mort s'esdevenia abans de la primera dècada en la vida del pacient.

Les mutacions que afecten a la proteïna inactiven la transducció de senyals relacionats amb la Btk (*Bruton's tyrosine kinase*). La Btk és una proteïna tirosina cinasa citoplasmàtica (PTK, *Protein Tyrosine Kinase*). Pertany a la família de Tec que contenen 5 dominis diferenciats. El primer és un domini d'homologia a pleckstrina (PH, *Plekstrin Homology*) que és típic de la família de cinases Tec. Aquest domini uneix inositols fosfat que permeten que tota la proteïna ancori a la membrana i pugui iniciar la transducció de senyal un cop ha estat activada. El següent domini és un domini Btk de només 26 aminoàcids que pràcticament només es troba en aquesta família de proteïnes i sempre després del domini PH. No és clara encara la funció d'aquest domini. Els tres dominis següents formen la configuració típica de les cinases citoplasmàtiques un domini d'homologia a Src 3 (SH3, *Src Homology 3*) seguit d'un domini d'homologia a Src 2 (SH2, *Src Homology 2*) i finalment el domini catalític de cinasa (Fig 6) (Mao et al., 2001).

La Btk pot ser activada per diferents senyals extracel·lulars, com pot ser el *crosslinking* de l'antigen receptor de la cèl·lula B o unió a interleucina-5 (IL-5). Després de la activació del receptor de cèl·lula B la Btk és fosforilada a la Tyr 551 per una tirosina cinasa de la familía Src i es transloca a la membrana plasmàtica. Posteriorment el domini SH3 sofreix una autofosforilació en la Tyr 223 que desencada l'activitat cinasa que transdueix la senyal a tota la via.

Tot i que s'han descrit diferents mutacions que afecten el gen i a la proteïna que condueixen a l'aparició d'agammaglobulinèmia, és d'interès destacar tres mutacions puntuals de les quals es coneix de manera més o menys clara el seu efecte sobre la proteïna.

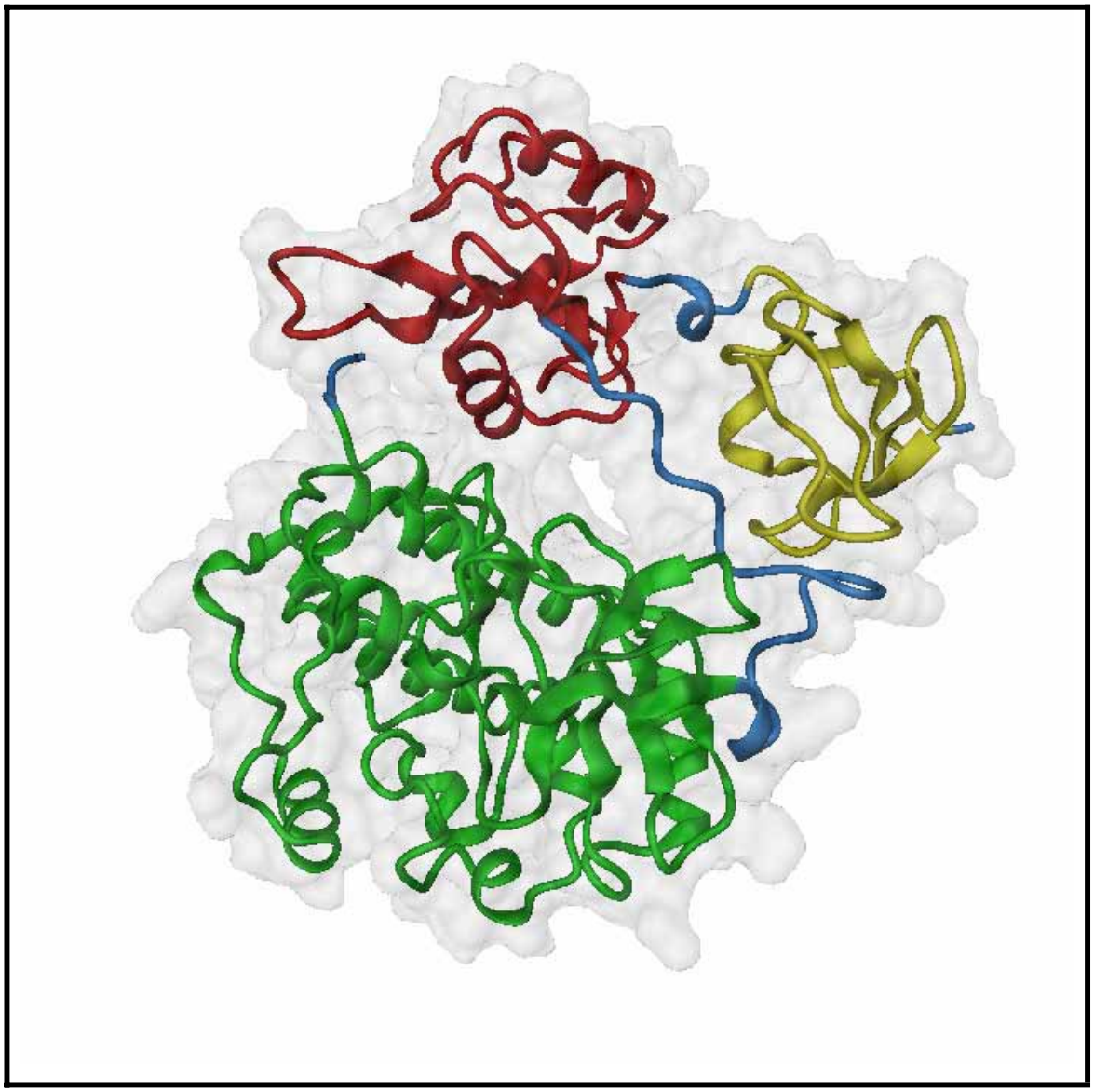

Fig 6. Estructura tridimensional de la Btk (part) (codi pdb 1K2P) En vermell domini SH3 en groc domini SH2 i en verd domini catalític cinasa.

Arg525Gln: La transició de G a A en el nucleòtid 1706 provoca el canvi en la proteïna de arginina a glutamina en la posició 525. Ja el 1993 (Vetrie et al., 1993) es va pronosticar que aquesta posició conservada havia de tenir un efecte

important en la funció catalítica de proteïna tirosina cinasa. L'efecte de la substitució de l'arginina impediria el reconeixement del substrat ja que es creu

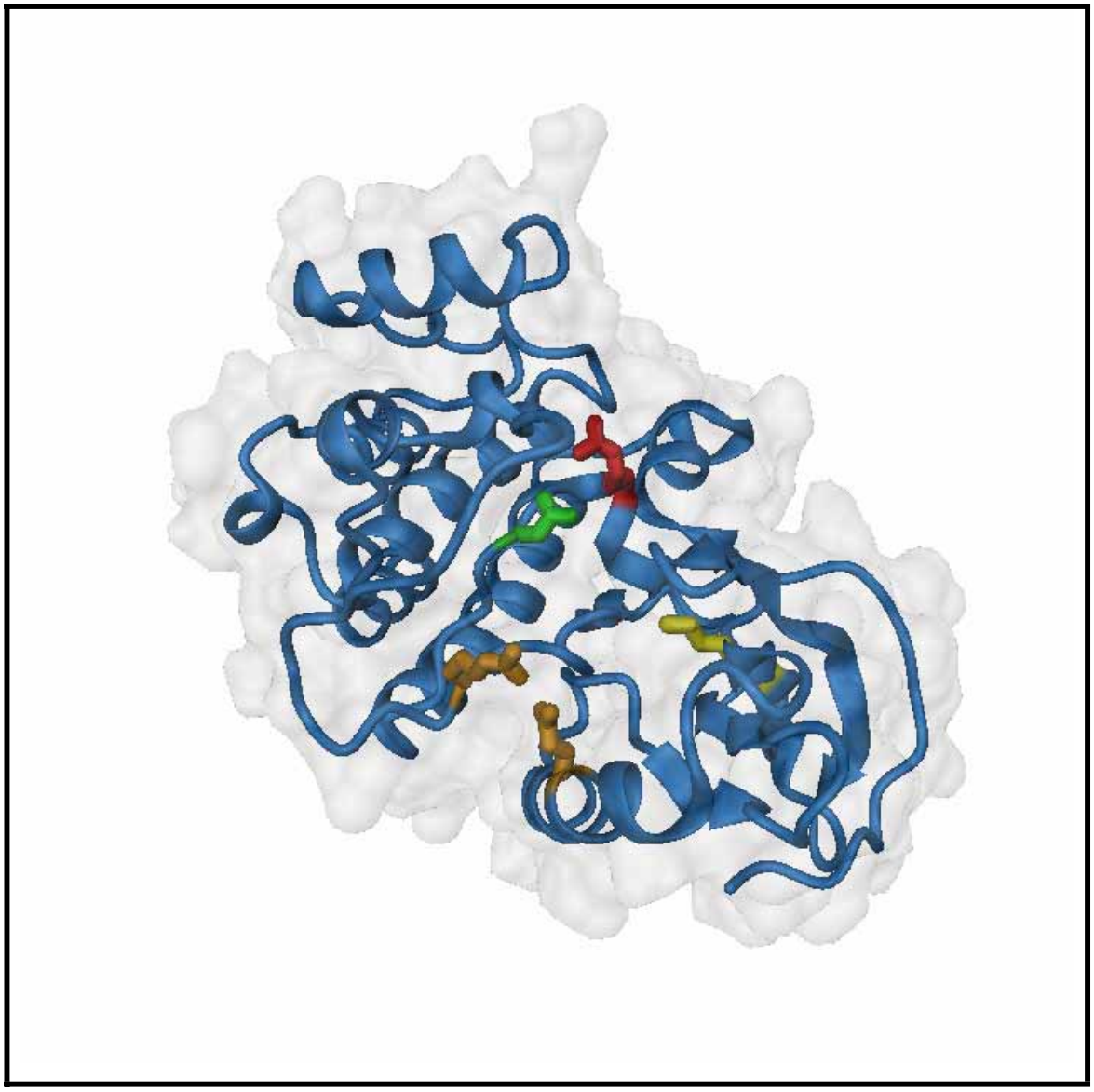

fig 7. Domini catalític de Btk. (codi pdb 1K2P). En verd Asp 521 residu catalític. En vermell Arg525, interacciona amb P-γ del ATP. En groc Lys 430 interacciona amb P-α del ATP.

que el residu és important en el domini d'unió del substrat. Posteriorment s'ha mostrat sobre l'estructura que aquest residu uneix l'ATP en la posició del fosfat γ i l'estabilitza durant la reacció de transferència que té lloc en l'aspàrtic situat en la posició 521 que és una posició invariant en la família PTK. L'aspàrtic activa per desprotonació l'hidroxil que serà responsable de l'atac nucleofílic sobre el

fosfat de l'ATP. En aquest cas concret l'hidroxil desprotonat serà el d'una tirosina (Mao et al., 2001) (figura 7). Aquesta mutació apart de ser causant de agammaglobulinèmia lligada al X també s'ha detectat en una família amb un diagnosi de immunodefieciència variable comú (McKusick-Nathans Institute for Genetic Medicine, 2000).

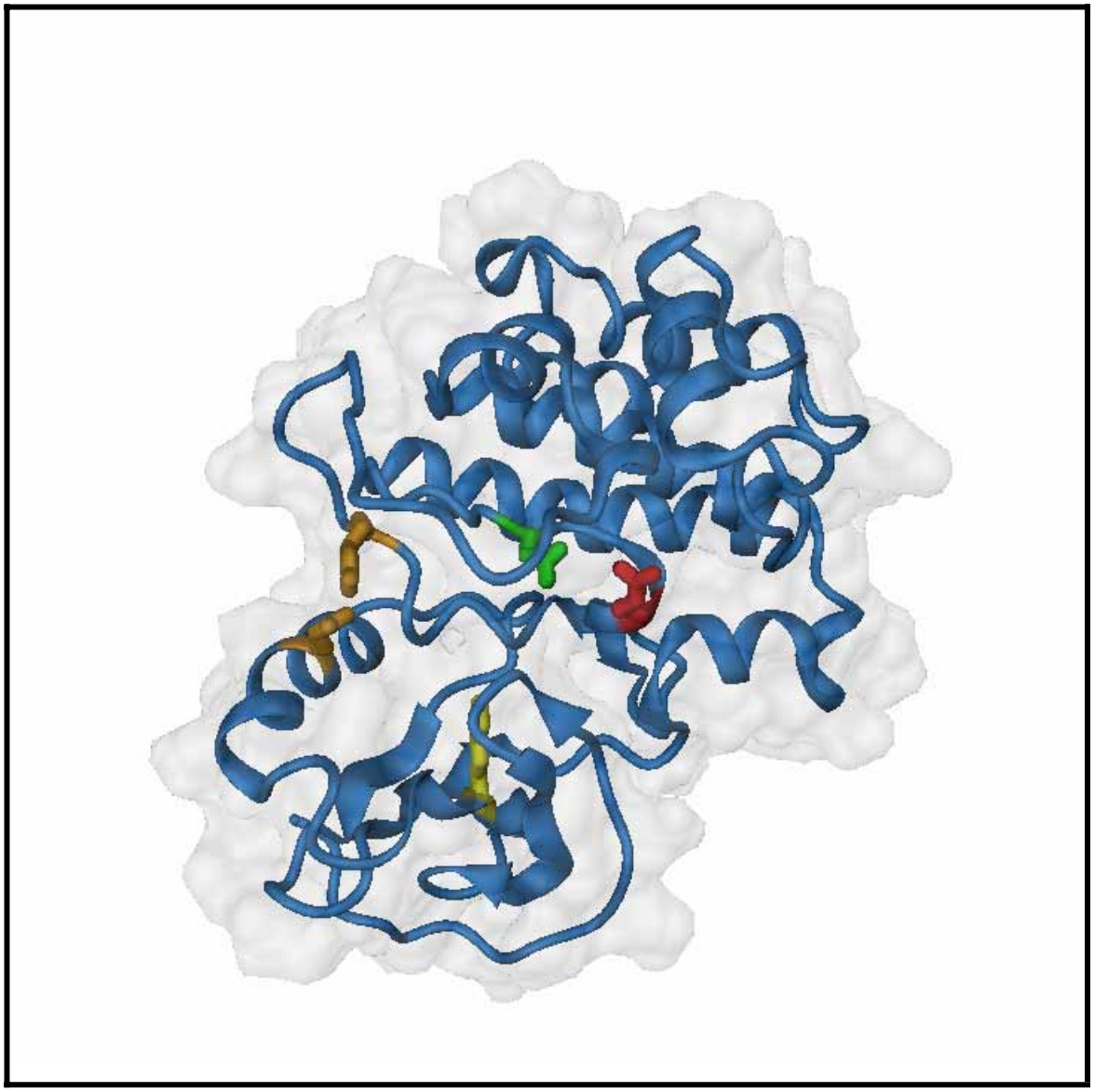

Fig 8. Domini catalític de Btk. (codi pdb 1K2P). En verd Asp 521 residu catalític. En vermell Arg525, interacciona amb P-γ del ATP. En groc Lys 430 interacciona amb P-α del ATP. En taronja, Glu445 sobre l'helix C i Arg 544, formen pont d'hidrogen que comporta una conformació no catalítica.

Lys430Glu: També s'ha identificat una mutació de transició de A a G en la posició 1420 que genera un canvi de lísina a glutàmic en la posició 430. Aquesta lisina que està molt conservada en tota la família de tirosines cinases i que es troba en el domini d'unió a l'ATP aboliria completament l'activitat cinasa. Segons Mao (Mao et al., 2001) i comparant l'estructura tridimensional obtinguda del domini catalític de la Btk amb altres estructures ja resoltes de PTKs s'observa

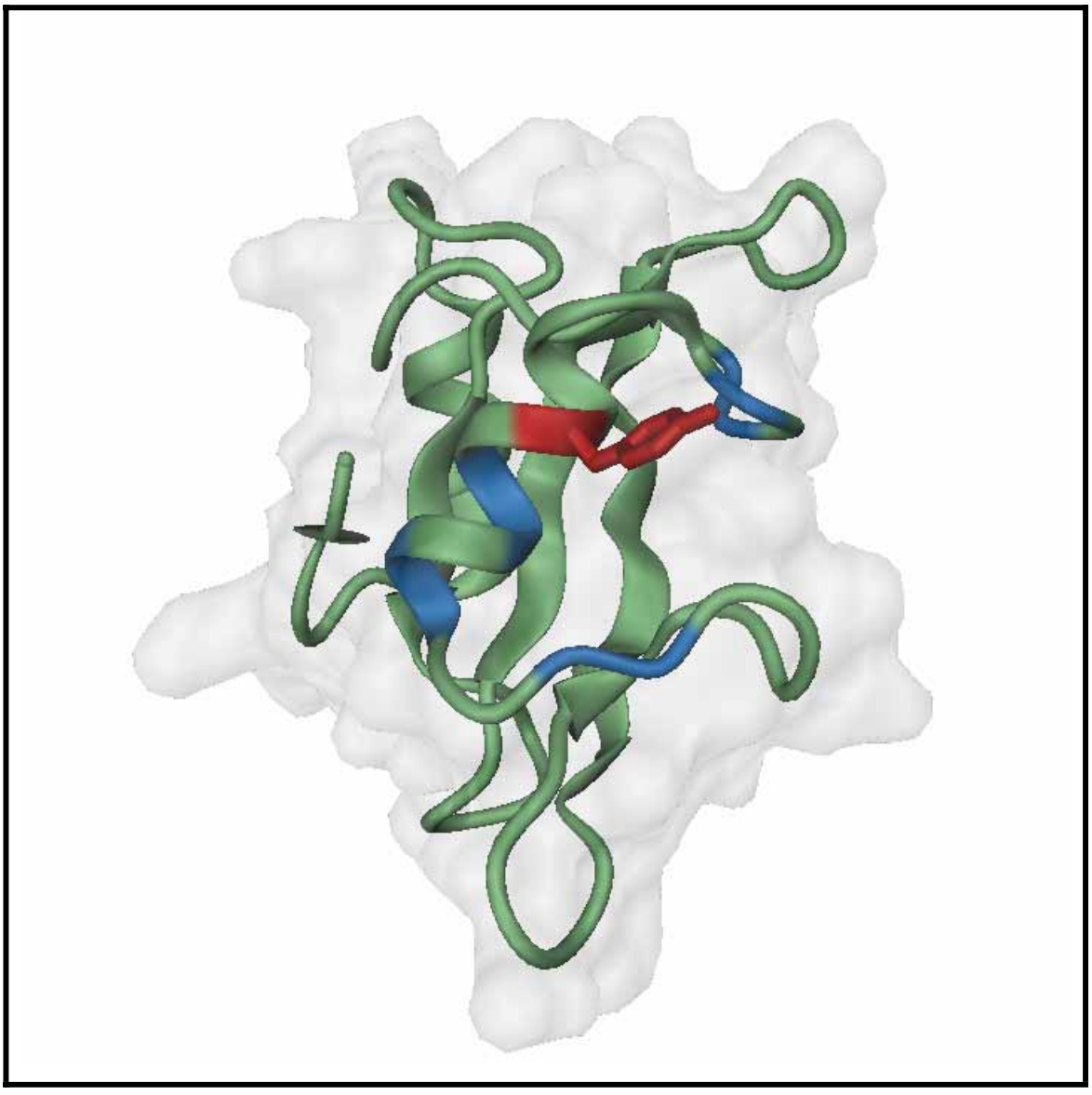

Fig 9. Domini SH2 de Btk (modelat usant modeller a partir del pdb 1QLP) En vermell Arg 361, en blau residus hidrofòbics veïns que formen la cavitat hidrofòbica.

que la unió del ATP ha de ser en els residus Lys430 amb P-α , Glu445 amb P-β aquest residu es troba sobre una hèlix que orienta tota la unió de l'ATP, i Arg525 amb P-γ, sent l'Aspàrtic 521 el residu catalític (veure figura 8). També sembla que s'ha de formar un pont d´hidrogen entre els residus 430 i 445 per la correcte orientació de la hèlix. Al cristall apareix un pont d'hidrogen entre el residu 445 i 544 que competeix amb la formació del pont d'hidrogen clau 430-445. Els autors hipotetitzen que si es fosforila la Tyr 551 fa que el glutàmic 544 interaccioni amb la tirosina fosforilada i per tant es perdi el pont d'hidrogen 445-544, el que permet la formació del pont d'hidrogen entre 430 i 445 que permet configuració adequada per a la unió amb l'ATP. Ara sembla clar que la mutació del residu 430 ha de comportar una invalidació de tot el mecanisme pel qual s'activa el sistema (McKusick-Nathans Institute for Genetic Medicine, 2000).

Tyr361Cys: Finalment s'ha trobat caràcter patològic al canvi de Tirosina a Cisteïna en cinases citoplasmàtiques com ho és la Bruton. Aquesta mutació cauria en una butxaca hidrofòbica del domini SH2, que juga un paper crític en la unió tal i com es veu a la figura 9. La conseqüència predita és un decreixement en l'estabilitat de la proteïna BTK, possiblement com a resultat de la inhabilitat de la BTK d'interaccionar amb els seus substrats.

-INSULINA

En aquests exemples un cas s'afecta a la interacció entre dues proteïnes hormona i receptor, en l'altre s'observa com l'alteració d'un residu impedeix el correcte reconeixement per la maduració de la proteïna i per tant no s'arriba a obtenir la forma funcional.

La insulina és una hormona clau en el metabolisme. És sintetitzada per les cèl·lules beta dels illots de Langerhans i consisteix en dues cadenes polipeptídiques diferents A i B, que estan unides per dos ponts disulfur. A

diferència de moltes altres proteïnes com la hemoglobina que està composada per subunitats estructuralment diferents, la insulina es troba sota el control d'un sol gen. Les dues cadenes A i B deriven d'una cadena precursora: la proinsulina. La proinsulina es converteix en insulina per un efecte enzimàtic que elimina el segment que uneix l'extrem amino-terminal de la cadena A amb l'extrem carboxi-terminal de la cadena B (veure figura 10). Aquest segment s'anomena C (de connector).

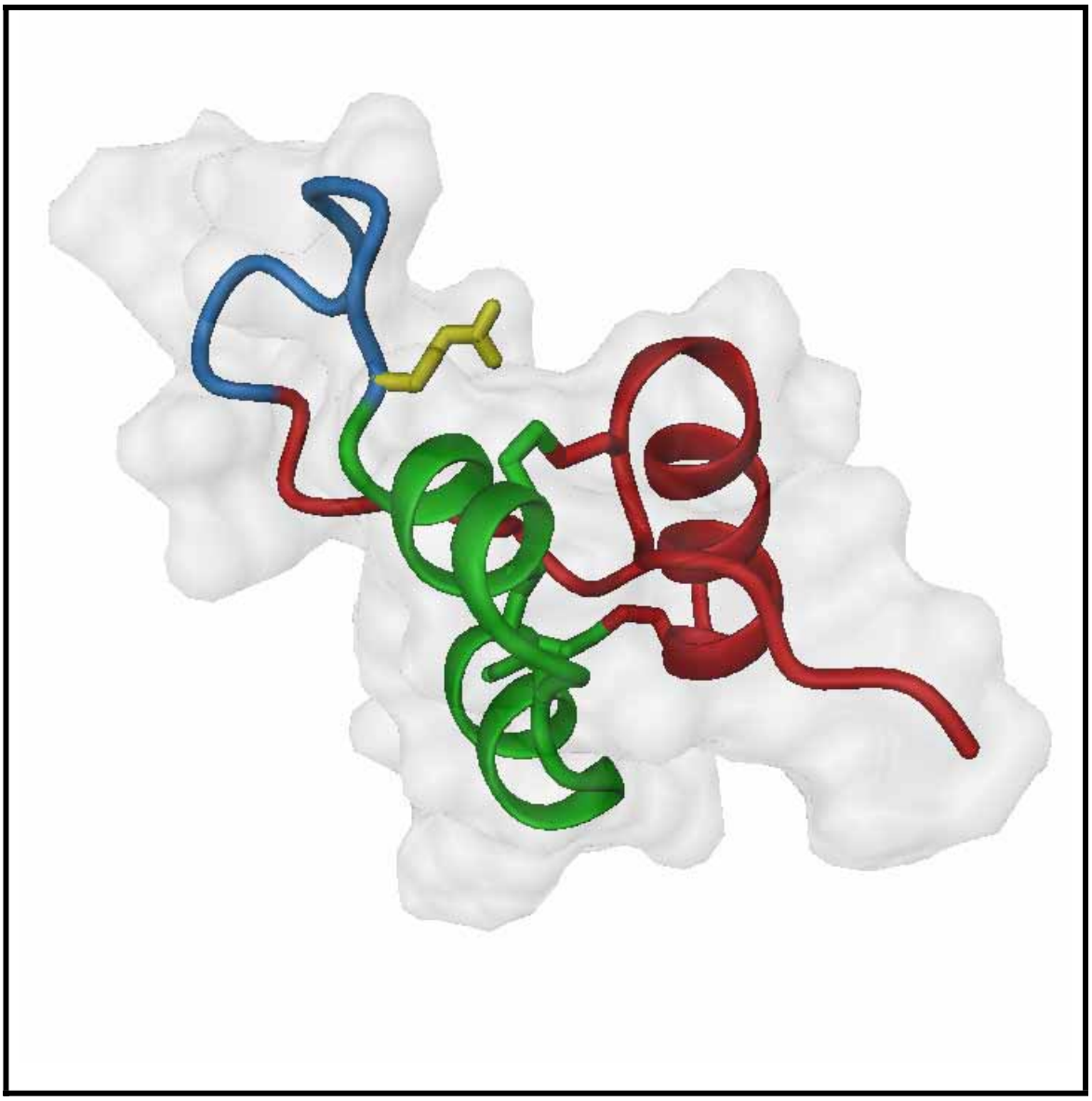

Fig 10. Estructura miniProinsulina (el pèptid connector, és més curt). La cadena B en verd, la cadena A en vermell i el pèptid C en blau. En groc residu Arg65 situat al pèptid C que inhibeix la maduració de la proinsulina.

Phe24Ser: La insulina conté tres exons, el segon codifica pel pèptid senyal la cadena B i part del pèptid C i l'exó tres codifica per la resta del pèptid C i la cadena A. En un malalt de diabetis mellitus de tipus II es va trobar una mutació en la posició 24 de la cadena B d'insulina, que generava un canvi Phe→Ser. El malalt presentava hiperglicèmia després d'ingestió sense mostrar resistència a l'administració exògena d'insulina. Posteriorment altres estudis en famílies de malalts es va concluir que mutacions en les posicions B24 i B25 s'associaven a diabetis mellitus de tipus II. i que la posició phe(B24) estava conservada a tots els mamífers. En els estudis sobre el cristall es va comprovar que l'anell aromàtic suportava tota la superfície d'interacció amb el receptor a través de interaccions a llarga distància d'empaquetament en el cor de la proteïna (figura 11). Finalment

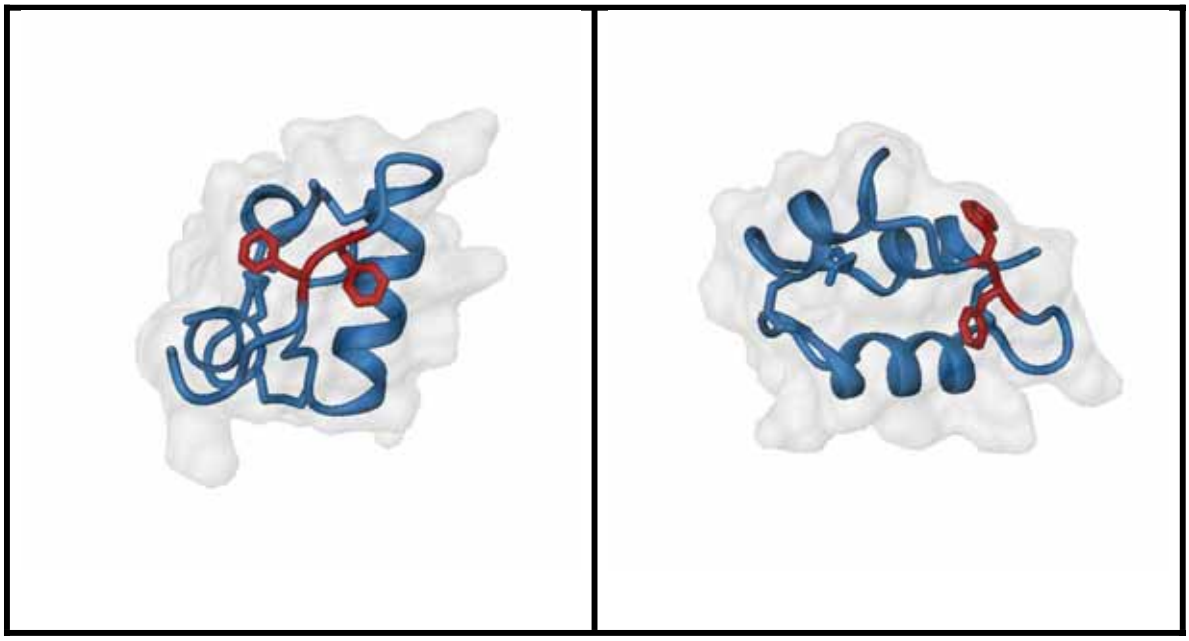

Fig 11. Estructura tridimensional insulina (codi pdb 1EFE). En vermell tirosines 24 i 25 en la cadena B. S'observen els ponts disulfur entre la cadena B i A.

es va trobar que l'anàleg gly(B24)-insulina presentava un desplegament parcial de l'estructura però a l'hora sorpresivament mantenia certa activitat biològica; també es va veure que el mutant ser(B24)-insulina, mantenia una estructura quasi-nativa però perdia gran part de l'activitat biològica. Això va fer sospitar als

científics que la insulina sofria un canvi conformacional al unir-se al receptor (Hua et al., 1993; McKusick-Nathans Institute for Genetic Medicine, 2000).

His65Arg: En l'anàlisi de una família japonesa amb hiperinsulinèmia es va trobar una mutació puntual que suposava una substitució d'histidina per arginina a la posició 65, el que evita el reconeixement de la proteasa dibàsica responsable de la maduració de la proinsulina a insulina (figura 10)(McKusick-Nathans Institute for Genetic Medicine, 2000).

VI. MUTACIONS PATOLÒGIQUES PUNTUALS: DESCRIPCIÓ

Com s'ha vist en els exemples anteriors el coneixement de les mutacions patològiques i el seu impacte en l'estructura/funció de les proteïnes té gran interès en la comprensió de les causes de la patologia. Pel seu estudi s'han seguit diferents aproximacions en els darrers anys. Alguns autors opten per estudiar mutacions humanes mentre que d'altres es decanten per models bacterians. El model que anomenarem humà es basa en recopilar en bases de dades mutacions específicament patològiques. Per complementar amb mutacions neutres s'usen dues estratègies diferents una complementària amb l'anterior que consisteix en cercar en les bases de dades humanes SNPs definits explícitament com a neutres, l'altra consisteix en generar un model neutre a partir de les proteïnes que són molt properes evolutivament a l'estudiada i agafant aquestes diferències com a mutacions neutres (l'anomenarem model evolutiu). Altres autors fan servir com a font de mutacions neutres i patològiques els estudis sobre mutagènesi massiva en proteïnes bacterianes o de virus. Aquests estudis assagen la funcionalitat de la proteïna *in vitro*. Els autors seleccionen les mutacions que afecten a la funció de la proteïna com a patològiques i les que no l'afecten com a neutre, deixant marge a les que afecten només parcialment per incloure-les o no en els grups de dades.

El que els autors intenten predir en aquests models és la funcionalitat de la proteïna. El número de proteïnes varia molt entre els dos esquemes ja que, el model bacterià es nutreix de tres proteïnes molt ben caracteritzades, mentre que el model humà usa desenes o centenars no tan ben definides.

Un dels primers treballs predictius sobre patogenicitat de mutacions apareguts és de Sunyaev et al (Sunyaev et al., 2000b) aquí els autors creen un grup de mutacions puntuals partint de la cerca de la base de dades SwissProt (Boeckmann et al., 2003) usant dues paraules clau, <estructura tridimensional> i <mutació patològica>. Així es recuperen mutacions puntuals associades a patologia en proteïnes humanes que tenen estructura tridimensional resolta. Es recullen fins a 551 mutacions puntuals patològiques. Per construir el grup control de mutacions puntuals neutres usen dues estratègies diferents. En primer lloc, de manera anàloga a les mutacions patològiques busquen mutacions puntuals no relacionades amb malalties, i enriquir el grup amb dades a la base de dades OMIM (McKusick-Nathans Institute for Genetic Medicine, 2000) variants al·lèliques amb una incidència en població superior o igual al 1%. Unint aquests dos grups de mutacions neutres obtenen 86 mutacions neutres. La segona estratègia emprada és la cerca a HSSP (Frishman & Argos, 1995) que és una base de dades d'alineaments múltiples de proteïnes que tenen la seva estructura resolta. Es cerquen proteïnes amb una identitat superior al 95% amb les proteïnes del grup de patològiques i consideren neutre la variabilitat inclosa en aquest 5% restant de disimilaritat. Així, s'obtenen 250 mutacions puntuals neutres. L'article de Sunyaev et al. és simple i preliminar però ja demostra l'existència de certes diferents entre patològiques i neutres. Així el 35% de les mutacions patològiques es troben enterrades i només el 9% de les neutres hi estan. També mostren que al voltant del 70% de les mutacions patològiques es troben en zones que són, amb alta probabilitat, estructuralment o funcionalment importants mentre que només el 17% de les neutres ho són.

En un treball posterior (Ramensky et al., 2002) els autors posen a punt un diagrama de treball per tal de calcular un seguit de paràmetres per cada mutació puntual. Determinen si la posició on es troba la mutació estudiada cau en alguna zona funcionalment important. També calculen la matriu de perfil, sobre els alineaments múltiples derivats de Blast (Altschul et al., 1997) usant els homòlegs que tenen una identitat de seqüència entre el 30-94%. La matriu de perfil és una mesura del grau de conservació en els alineaments múltiples que correspon al logaritme de la raó entre la probabilitat d'un aminoàcid d'ocórrer en la posició de l'alineament estudiada, sobre la probabilitat de l'aminoàcid d'ocórrer en qualsevol posició. Obtenen els valors pel residu salvatge i pel mutant. També mapen la mutació sobre l'estructura, a partir de la qual s'obtenen, valors d'accessibilitat a solvent relativa, la zona dins el mapa de Ramachandran segons els angles dihedres φ-ψ, i el factor B de l'estructura cristall pel residu i la pèrdua de ponts d'hidrogen. Utilitzen també canvis en les propietats dels aminoàcids involucrats, com són canvis en el potencial hidrofòbic i en el volum degut al canvi observat. Finalment també analitzen contactes dels residus amb lligands, amb altres subunitats i amb residus molt conservats en els alineaments.

En el treball desenvolupat per Wang i Moult (Wang & Moult, 2001) el grup de mutacions patològiques es genera partint de tots els SNPs que estan a la *dbSNP database* (http://www.ncbi.nlm.nih.gov/SNP), d'aquests seleccionen els que es troben dins proteïnes i són no-sinònims. D'aquests es cerca quins SNPs són patològics usant de nou la base de dades OMIM (McKusick-Nathans Institute for Genetic Medicine, 2000). Aleshores busquen per quines proteïnes hi ha estructura tridimensional o bé algun homòleg amb una identitat de seqüència superior al 25%. Al final 157 proteïnes són seleccionades amb almenys una mutació puntual patològica. D'aquestes els autors en seleccionen només 23 aleatòriament per fer-ne l'estudi. El grup de mutacions neutres és generat usant 150 gens relacionats amb la homeòstasi de la pressió sanguínia i extrets del treball de Halushka (Halushka, 1999) i 106 gens relacionats amb malalties

cardiovasculars endocrinologia i neuropsiquiatria obtinguts del trebal de Cargill (Cargill et al., 1999). Amb els dos grups de mutacions els autors construeixen fins a 5 regles per assignar els efectes de les mutacions. Aquestes cinc regles incideixen en l'estabilitat proteica, en relació a pèrdua de ponts d'hidrogen, d'interaccions hidrofòbiques, d'aparició de càrregues en zones enterrades, introducció de volums grans en zones enterrades entre altres. Les quatre regles següents són més simples, intenten respondre si s'afecten residus funcionalment importants pel que fa a la unió de lligands, catàlisi, regulació al·lostèrica i modificacions post-traduccionals. Els resultats dels autors es basen en classificar les diferents mutacions segons el compliment o no de les regles. Així mostren que el 90% de les mutacions patològiques compleixen alguna de les regles i més del 80% ho fan amb la primera que és la de l'estabilitat proteica. Només el 5% de les mutacions patològiques afecten centres actius o importants funcionalment. El 10% de les mutacions patològiques no són classificades com a tal usant les regles establertes. De l'anàlisi de les mutacions neutres s'extreu que el 70% no compleixen cap de les regles establertes. Del 30% restant que en compleixen alguna, la gran majoria compleix la regla que refereix a l'estabilitat de la proteïna. L'explicació dels autors a aquest fet rau diverses raons, l'error en l'assignació de mutacions, o funcions que són compensades per altres mecanismes dins la cèl·lula.

Un altre treball interessant correspon a Ng i Henikoff (Ng & Henikoff, 2001), on presenten un nou mètode per tal de predir mutacions puntuals patològiques basant-se en càlcul de probabilitats normalitzades de tots els canvis possibles per una posició al llarg de tota la columna de l'alineament. Deriven directament les probabilitats associant-les a les freqüències relatives de cada aminoàcid que apareix en cada columna de l'alineament. Apliquen un llindar a les probabilitats calculades, per sota del qual una mutació serà patològica. Aquest llindar és que la probabilitat normalitzada de l'aminoàcid patològic sigui inferior a 0.05 en la columna de l'alineament de la posició de la mutació. Anomenen el mètode SIFT

(*sort intolerant from tolerant*). La idea una vegada més consisteix en recollir homòlegs per una proteïna usant Blast i alinear-los i calcular-ne les probabilitats normalitzades per tots els canvis possibles en cada posició de l'alineament. En la fase d'assaig del mètode els autors usen una nova aproximació que consisteix en usar dades de mutagènesi exhaustiva sobre tres proteïnes model, LacI (anomenem grup LAC), la proteasa de HIV-1 (HIV) i el lisozim del bacteriofag T4 (T4). Per aquestes tres proteïnes hi ha assaigs de funcionalitat sobre una bateria bastant exhaustiva de mutacions puntuals. Els autors seleccionen les mutacions d'aquests assaigs en dues categories. Les mutacions que causen una pèrdua de funció són aquelles que l'activitat de la funció de la proteïna es troba disminuïda o abolida, neutres aquelles que no afecten o molt suaument a la funció. També es descarten les mutacions que són poc severes. Amb aquests grups de dades generats, optimitzen els llindars dels valors de SIFT i per la matriu de substitució BLOSUM62 (Henikoff & Henikoff, 1993) per tal de maximitzar la diferència entre neutres i les de pèrdua de funció. Aquesta optimització la fan sobre el grup de dades de LAC; després amb els llindars determinats ho apliquen a HIV i T4 per tal de demostrar que el mètode es pot generalitzar a altres proteïnes. Els resultats obtinguts sobre el grup de LAC que són un total de 4004 mutacions puntuals el mètode reconeix com a neutres el 78% de les mutacions que són realment neutres o tolerants mentre que reconeix com a pèrdua de funció el 57% de les mutacions que perden la funció o es veu disminuïda dràsticament. En total, sobre totes les mutacions classifica correctament el 68% de les mutacions. El comportament usant BLOSUM62 decreix notablement malgrat que continua tenint una capacitat discriminatòria. Els resultats pels altres grups de dades és similar encara que el número total de mutacions és sensiblement inferior ja que són 336 per HIV i 2015 per T4. La pèrdua de capacitat predictiva de BLOSUM en front la matriu de substitució específica de posició s'explica segons els autors pel fet que BLOSUM recull

informació de molts alineaments diferents mentre que SIFT recull informació que és pròpia de cada seqüència i el seu alineament.

Chasman i Adams mostren en el seu treball (Chasman & Adams, 2001) que l'ús de propietats tridimensionals i evolutives en l'anàlisi de les mutacions puntuals aporta informació molt útil per tal de classificar les mutacions en neutres i patològiques. Aquests autors usen com a grups de dades dos dels grups usats pels autors Ng i Henikoff esmentats previament, LAC i T4 i en calculen diferents paràmetres. Alguns tenen valors continus com accessibilitat a solvent, entropia filogenètica relativa, factor-B de l'entorn tridimensional. Altres tenen valors categoritzats com el nombre de vegades que surt l'aminoàcid mutant en la columna de l'alineament, càrregues enterrades, alteracions d'elements d'estructura secundària, proximitat a lligands o interfícies entre altres.

Amb tots aquests paràmetres determinats per totes les mutacions els autors apliquen l'estadística per determinar si les distribucions de valors per neutres i patològiques són significativament diferents.

Ja posteriorment Steward et al. (Steward et al., 2003) presenten una anàlisi més exhaustiva de mutacions patològiques derivades de la base de dades OMIM. El grup de dades de mutacions patològiques arriba 5686. Les mutacions neutres (1581) són obtingudes de la base de dades dbSNP. El primer que mostren és que les proteïnes relacionades amb malaltia segueixen una distribució de funcions segons les anotacions de *gene ontology* (GO) similar a la del proteoma humà al complet. D'aquesta manera l'autor mostra que les proteïnes humanes que són relacionades amb malaltia no configuren un subgrup dins el proteoma humà sinó que en general poden estar relacionades amb qualsevol tipus de funció. Una altra anàlisi següent consisteix en escollir les mutacions que tenen estructura coneguda i mapar-ne les mutacions. En total són 63 proteïnes i unes 1300 mutacions patològiques. Analitzen el grau de conservació i l'accessibilitat de les mutacions respecte a la resta de posicions conegudes. Mostren que s'acumulen mutacions patològiques en zones amb un alt grau de conservació i en zones enterrades però

no hi ha grans diferències respecte el que seria el rerefons entès com la resta de mutacions puntuals. Finalment fan una anàlisi molt detallada d'alguns exemples basats en les malalties que causen i en l'efecte de la mutació sobre l'estructura. Així avaluen efectes de canvis de volum en zones interiors i empaquetades de proteïnes, introducció de càrregues en zones enterrades, alteració d'interaccions de proteïna-proteïna, alteracions de xarxes d'enllaços d'hidrogen, alteracions en l'unió a ADN, trencament de ponts de sofre i mutacions en centres actius o d'unió.

Caldria destacar també destacar també el treball de Vitkup et al. (Vitkup et al., 2003) aparegut el 2003. Els autors parteixen d'un grup de dades de mutacions puntuals patològiques extretes de la base de dades OMIM i referenciades a SwissProt, en total són 4236 mutacions que afecten a 436 gens. Com a mutacions neutres apareixen 1037 SNPs sinònims i no sinònims que provenen de l'anàlisi massiva d'haplotips en 313 gens. Representen tots aquests canvis en dues matrius de substitució on cada canvi és comptat en la cel·la que li correspon. Aquestes dues matrius les comparen també amb dues matrius més, una que anomenen de freqüències esperades i basada en freqüència de codons i l'altra que anomenen interespècie i que està basada en la matriu evolutiva PAM1. La matriu esperada representaria les freqüències de transicions aminoacídiques en absència de selecció. La matriu interespècie correspon a la matriu PAM1 de Dayhoff que correspon a distàncies evolutives molt curtes i normalitzada per la freqüència d'aminoàcids del genoma humà. Les matrius patològica i neutre tenen molts elements amb valors iguals o propers a 0 i no són simètriques, les tendències als canvis de *i* a *j* són diferents a les tendències de *j* a *i*. La interpretació d'aquestes matrius esdevé complexa i confusa sense obtenir grans tendències, però en general les seves conclusions, són concordants amb la resta de resultats anteriors.

VII. MUTACIONS PATOLÒGIQUES PUNTUALS: PREDICCIÓ

En els treballs estudiats en l'apartat anterior observem que existeix una sèrie de variables, relacionades amb l'estructura de la proteïna, que ens permet distingir entre mutacions patològiques i neutres. Això planteja la possibilitat de predir quan una mutació serà patològica usant aquestes variables. En els propers treballs intenten establir ja metodologies per predir l'efecte sobre la salut de SNPs al llarg del genoma intentant aplicar metodologies que ja han estat descrites en l'apartat anterior.

En el treball presentat de Sunyaev (Sunyaev et al., 2000b) presenta a partir de totes les dades derivades de la descripció una metodologia encarada a la predicció del caràcter patològic de les mutacions puntuals. Amb tots els paràmetres calculats deriven regles empíriques per determinar si un SNP serà o no patològic. Les regles són 8 i combinen 3 característiques diferents: valor de la matriu de perfil, propietats derivades del lloc de la mutació i propietats derivades de la substitució. Si les tres regles es compleixen la classificació diu que la mutació serà neutre o probablement patològica o possiblement patològica. En aquestes regles finals s'exclouen molts dels paràmetres calculats ja que són redundants.

Els autors mostren que pel 82 % de les mutacions patològiques de la base de dades SwissProt són predites correctament usant les seves regles i només tenen un 8% de falsos positius usant com a grup de mutacions neutres el model derivat dels alineaments múltiples. Amb aquestes dades intenten predir els SNPs de la base de dades HGVbase. Sobre un total de 11152 SNP no-sinònims 1591 són predits com a possiblement patològics i 1257 més com a probables usant només dades derivades de seqüència.

En el treball de Sunyaev et al. (Sunyaev et al., 2001) s'apliquen les regles que havien anat mostrant en treballs anteriors per tal de predir polimorfismes de diferents bases de dades. Apliquen diferents regles que tenen en compte paràmetres estructurals així com evolutius. Determinen el seu grau d'error que es troba al voltant del 10-30% per les mutacions patològiques, que són 1550 mutacions puntuals identificades en pacients i extretes de la base de dades SwissProt. El percentatge d'error per les mutacions neutres són del 9% i estan derivades de les diferències entre les proteïnes humanes i les seves ortòlogues en altres espècies però molt properes en l'evolució (és a dir més del 95% d'identitat de seqüència). Aquests resultats els fan servir per predir SNPs de les bases de dades públiques Els seus resultats mostren que un 30% dels 245 SNPs ben caracteritzats són patològics o més concretament són predits com que danyarien l'estructura o bé la funció. Els autors ponderen que si la seva taxa de falsos positius és del 9% aleshores un 20% dels SNPs que es troben a la base de dades públiques serien patològics. En una anàlisi més detallada de 99 SNPs no sinónims que tenen estructura tridimensional coneguda i també la seva freqüència al·lèlica. A més a més, d'aquests 99 SNPs en 11 casos se sap que tenen associació amb malaltia de manera clara. D'aquests 11, 8 SNPs (73%) són predits com a patològics. Amb aquesta predicció sobre dades d'alta qualitat se'n deriva una estimació global sobre tot el genoma humà. Segons aquesta estimació cada individu seria portador d'unes 2000 variants que afecten a la funció o l'estructura de les proteïnes. També Ng i Henikoff (Ng & Henikoff, 2002) presenten l'aplicació del seu mètode SIFT a la predicció de SNPs de les bases de dades públiques. Els seus resultats mostren fins un 69% d'encert en mutacions patològiques derivades de la base de dades SwissProt i un 81% d'encert en la predicció de les mutacions neutres derivades de la base de dades de SNPs neutres del Whitehead Institute (Cargill et al., 1999). El mètode és posa a prova amb un grup de dades independent. S'usen els SNPs no codificants de la base de dades dbSNP. Donat que el mètode necessita de la presència d'homòlegs per l'anàlisi

només poden analitzar el 60% de les proteïnes de la base de dades i que representen uns 3000 SNPs codificants no sinònims. D'aquests un 25% són predits com a patològics i la taxa de falsos positius fa que siguin un 19% dels SNPs de la base de dades patològics. Els autors fan un estudi detallat de 16 proteïnes que tenen un número alt de SNPs predits com a patològics. Mostren que moltes mutacions predites com a patològiques realment han estat identificades en pacients. Discuteixen també el fet que la predicció com a pèrdua de funció pot resultar en un fenotip neutre ja que la funció de la proteïna pot ser substituïda per un homòleg o bé la funció pot no ser clau per la supervivència de l'individu. Els autors també detecten errors en la identificació i mapat dels SNPs que introdueix soroll en les bases de dades. Els errors poden venir de la seqüenciació així com de la confusió entre gens i pseudogens. Es mostra també l'avantatge del mètode SIFT sobre altres mètodes ja que no necessita d'estructura.

En el treball de Martin et al (Martin et al., 2002)els autors intenten recollir tot el que s'ha fet en el camp de la caracterització de mutacions patològiques i ho aplica a una sola proteïna d'alt interès biomèdic com és la proteïna supressora de tumors p53. A diferència dels treballs anteriors, que eren molt més generalistes, en aquest treball els autors usen una sola estructura de la qual es coneixen fins a 882 mutacions i intenten aplicar tot el que es coneix per tal de donar explicacions estructurals i evolutives a cada una d'elles. Per cada mutació determinen un paràmetre relacionat amb la variabilitat de la posició en l'alineament múltiple. A nivell estructural analitzen la pèrdua de ponts d'hidrogen així com la formació de l'efecte estèric de les mutacions observades sobre l'estructura de la proteïna. També analitzen si la mutació té lloc en residus relacionats amb l'interacció amb el ADN, amb la unió a zinc, i les mutacions a prolina i mutacions desde glicina. Així usant criteris només estructurals expliquen el 35% de les mutacions, incloent informació evolutiva expliquen fins els 56% de les mutacions. És interessant de ressaltar que el 65% de les mutacions no explicades es troben a la

superfície i per tant podrien ser explicades per alteracions en les interaccions de la proteïna amb altres proteïnes o altres dominis de la pròpia.

En el treball de Chasman i Adams (Chasman & Adams, 2001) també es presenta una extensa anàlisi de la capacitat de predicció del seu mètode. En general mostren que són més discriminats els paràmetres continus que els categòrics i que la distribució de valors és essencialment la mateixa pels dos grups de mutacions LAC i T4. Els paràmetres amb més poder discriminant són accessibilitat relativa i el factor B que són valors exclusivament estructurals. Dins els paràmetres categoritzats els més informatius són el de que la mutació introdueix una càrrega en una posició enterrada i la següent és la presència d'un aminoàcid inusual en la columna de l'alineament. S'introdueix un nou concepte que és el de la probabilitat que una mutació afecti o no a la funció. A diferència dels treballs anteriors en què només s'establia un criteri o un punt de tall a partir del qual una mutació passava a afectar la funció o era patològica, ara els autors li donen un valor de probabilitat. Ho raonen pel fet que mutacions que estiguin molt properes al punt de tall podrien ser classificades com que afecten a la funció i en realitat no ho fan i al revés. La predicció es fa segons la probabilitat, si és 0.5 o superior afectarà a la funció amb un nivell de confiança igual al valor de la probabilitat. Per contra un mutació no afectarà a la funció si la probabilitat es inferior a 0.5 i el nivell de confiança serà igual a 1 menys la probabilitat. Ara les prediccions poden ser jutjades pel seu nivell de confiança.

Per primera vegada s'introdueixen dos procediments de validació creuada per tal de donar més versemblança a les seves prediccions. La *validació creuada i homogènia* consisteix en dividir cada un dels grups de dades LAC i T4 per separat, s'usa una meitat per entrenar i l'altre per predir així obtenen uns rendiments de predicció que són dependents de cada proteïna. En aquest cas el subgrup d'entrenament conté el 90% de les mutacions triades a l'atzar i el subgrup de predicció conté el 10% restant. Els paràmetres usats en cada predicció varia segons el grup de dades estudiat i s'escull pel mètode de l'aproximació de

màxima probabilitat d'entre la resta de paràmetres, és a dir s'usen els que tenen més poder de predicció per si sols. Obtenen un percentatge d'error al voltant del 25% quan usen un nivell de confiança del 0.5 i un error del 0.05% quan usen un nivell de confiança del 0.95 La segona validació és *heterogènia* i és més exigent ja que s'entrena amb una proteïna i es prediu en l'altre i al revés. En aquesta situació l'exactitud del mètode és menor però encara bona, al voltant el 30-35% amb un nivell de confiança del 0.5.

Finalment els autors intenten predir SNPs de dues bases de dades diferents, la del Case Western Reserve University i la del Whitehead Institute. Mapen automàticament els SNPs sobre estructures i intenten predir l'efecte usant com a grup d'entrenament els grups de dades de T4 i LAC. Es presenten dos problemàtiques, trobar estructures on mapar les mutacions que són poques al voltant del 30% i que poden incrementar fins al 50% de les mutacions usant models. La segona problemàtica és que moltes mutacions no tenen prou mutacions en el grup d'entrenament amb paràmetres similars com per poder fer prediccions amb rigor estadístic. Malgrat això arriben a la conclusió que al voltant del 30% dels SNPs en les bases de dades analitzades afectarien a la funció. Segons els autors aquestes xifres són més baixes que el percentatge observat en T4 i LAC que correspon al voltant de 40-44%. Extrapolen que si en les bases de dades que ells analitzen de SNPs estimen entre 24000 i 40000 locis heterozigotics per persona de mitja, d'aquests entre 6200-12800 afectarien a la funció de la proteïna.

Un pas més en aquest anàlisi és fet per Saunders i Baker (Saunders & Baker, 2002), en aquest treball els autors intenten clarificar tota la feina feta pels autors anteriors i intenten establir quins són els paràmetres més informatius en la classificació de SNPs. Usaran els dos grups diferents de dades fins ara usats, per una banda usen els derivats de mutagènesi dirigida i anàlisi funcional, que són LAC ,T4 i HIV. També usen un seguit de dades derivades de SwissProt de mutacions patològiques en humans i variabilitat neutre. A diferència dels autors

anteriors són molt més restrictius en el tipus de mutants que seleccionen. Així per les tres proteïnes seleccionen només aquelles mutacions que són molt lesives com pèrdua de funció i les que són molt neutres per les que no afecten a la funció. D'aquesta manera al final tenen fins 1500 mutacions que afecten a la funció i fins 3700 neutres. El grup de dades derivat de SwissProt és molt més petit i conté 191 mutacions patològiques i 87 neutres, els mateixos autors dubten de la validesa d'aquest grup de dades ja que està molt esbiaixat i perquè les neutres no és clar que hagin estat prou testades com a neutres. El primer anàlisi portat a terme és l'avaluació de la capacitat de classificació segons els diferents paràmetres usats. Aquests autors usen els paràmetres establerts per altres treballs i els avaluen usant una validació creuada, dividint les dades en dos grups a l'atzar un d'entrenament i un de predicció. Aquesta validació creuada es fa 20 vegades i s'obté el valor mitjà. Obtenen errors de la classificació per neutres i per patològiques i el resultat global balancejat ja que el número total de neutres i patològiques no és el mateix. Els paràmetres avaluats són de nou uns que depenen de la seqüència i uns que depenen de l'estructura. Per seqüència tenim: valors de matriu de substitució BLOSUM62, Entropia normalitzada per lloc, SIFT que correspon a la matriu de perfil implementada per Ng i Hennikoff(Ng & Henikoff, 2002) i comentada anteriorment. Els paràmetres que depenen de l'estructura són: densitat de carbonis β, Accessibilitat a solvent, factor B normalitzat i les regles presentades per Sunyaev et al.(Ramensky et al., 2002). Usant només un d'aquests paràmetres és veu que el SIFT té un poder discriminatori més gran ja que l'error de classificació balancejat es d'un 22% i correspon a un 20% en patològiques i un 24% en neutres. Aquest resultat és més dolent que el mostrat pels propis autors de SIFT que no realitzaven validació creuada. L'accessibilitat al solvent i els factors B normalitzats prediuen molt bé les mutacions patològiques però no funcionen tant bé per predir mutacions neutres. Les regles de Sunyaev funcionen molt bé per neutres però fallen molt per patològiques. Els resultats pel grup de dades humanes derivades de la base de

dades SwissProt presenta resultats lleugerament pitjors. Seguidament els autors intenten combinar diferents paràmetres per tal d'augmentar el poder predictiu i es queden com millor combinació SIFT i la densitat de Carbonis β que dóna un percentatge d'error balancejat del 20% per les 3 proteïnes i un 29% per les mutacions humanes. Finalment intenten classificar les mutacions humanes amb els paràmetres derivats de les mutacions en proteïnes i observen que els resultats es mantenen al predir en un grup de dades totalment independent.

En el treball de Vitkup (Vitkup et al., 2003) també presenta un apartat de predicció a partir de la descripció presentada del model bacterià. L'usen per predir l'efecte de SNPs humans malgrat que el mètode ha estat desenvolupat per proteïnes evolutivament molt allunyades de l'espècie humana.

De manera similar Santibáñez-Koref (Santibanez-Koref et al., 2003) també usa la proteïna p53 i una metodologia similar per tal de classificar les mutacions patològiques. Fins a 20 paràmetres diferents són calculats per les mutacions patològiques, la majoria són paràmetres físico-químics com poden ser caràcter hidrofòbic, punt isoelèctric, volum, tendències a formar hèlix alfa, cadena beta entre altres. També determinen dos valors evolutius derivats dels alineaments múltiples, si la posició és conservada en tot l'alineament i si el mutant és present en alguna altra espècie alineada. Són mutacions neutres aquells canvis observats en altres espècies. Per cada mutació en deriven un *z-score* i poden comparar distribucions dels valors de *z-score* per mutants i neutres. És d'interès destacar l'anàlisi evolutiu que fan a partir de l'arbre filogenètic derivat de l'alineament múltiple on en poden derivar les taxes de mutació i el punt de divergència usant metodologies de l'anàlisi filogenètica.

Una aplicació semblant la trobem al treball de Mirkovic (Mirkovic et al., 2004) on intenten usar paràmetres estructurals per explicar mutacions puntuals mapades en la proteïna BRCA1 que està molt relacionada amb predisposició a càncer de mama i d'ovaris. Es parteix de 94 mutacions puntuals de la proteïna de les quals se sap si són neutres o provoquen algun canvi en la funció o en l'estructura que

els associa a algun càncer. Els autors deriven una vegada més un seguit de paràmetres estructurals, físico-químics i evolutius. També mapen sobre l'estructura cristall tots els mutants i addicionalment en generen un model per cada mutant. Per totes les mutacions intenten racionalitzar les causes per les quals apareix efecte patològic o no. D'aquesta manera expliquen amb raons estructurals, físico-químiques i evolutives 32 de les 37 mutacions patològiques. D'aquesta manera poden generar un seguit de regles interpretatives que queden plasmades en forma d'arbre de decisió. Aquest arbre implementat en forma de pàgina web és usada per predir 57 mutacions sobre les quals no es té cap informació sobre el seu efecte. 32 mutants són predits com a patològics o almenys generen un efecte sobre l'estructura o la funció. La resta són predits com a benignes o neutres.

En aquest apartat es mostren dues tendències diferents, la primera representada per Ng i Henikoff (Ng & Henikoff, 2002) i per Sunyaev et al. (Sunyaev et al., 2001) en la que intenta aplicar els mètodes descrits en l'apartat anterior per predir dades ja provinents de projectes de mapat de SNPs a gran escala. Els autors intenten confirmar les metodologies emprades usant ja milers de mutacions provinents de centenars de proteïnes humanes de les bases de dades i en general els rendiments de les prediccions són similars als presentats en treballs anteriors. Els resultats de predir sobre dades noves mostren que segons els autors entre un 15 i un 30 % dels SNPs tindrien un cert caràcter patològic o almenys alterarien la funció. Una vegada més el rigor estadístic no és molt gran i no existeix cap mena de validació creuada de les dades.

La segona tendència representada pels treballs de Martin et al. (Martin et al., 2002), de Mirkovic et al.(Mirkovic et al., 2004) i de Santibàñez-Koref (Santibanez-Koref et al., 2003) se centra en predir a les mutacions que apareixen en proteïnes d'interès biomèdic. Aquestes proteïnes, que són molt estudiades acumulen un conjunt elevat de mutacions puntuals patològiques i els autors apliquen paràmetres ja descrits i de nous per tal de predir i donar explicació a

l'efecte de la mutació. De manera general s'expliquen de manera satisfactòria un nombre elevat de casos però malgrat tot encara resten mutacions no explicades que podrien estar relacionades amb interaccions amb altres proteïnes.

A la taula 2, es mostra un petit resum dels mètodes explicats on es ressalta el número de proteïnes i mutacions, l'origen d'aquestes així com la metodologia en què es basa la predicció. En ella es veu clarament les dues estratègies explicades així com la variació entre autors del número de variants. Cal destacar també que molts autors no fan servir un mètode de validació creuada, el que disminueix la confiança en les dades de merit. Pel que fa als paràmetres més informatius hi ha dos grans grups de paràmetres, els estructurals i els evolutius. En general mutacions que alteren en alguna manera l'estabilitat de la proteïna tenen tendència a ser patològics. La manera com es mostra aquesta desestabilització varia entre els autors, alguns ho mostren segons el grau d'enterrament de les mutacions, altres com a trencament o alteració de l'entorn físico-químic. Els paràmetres evolutius exploren el grau de conservació de la posició de la mutació o en quina freqüència apareix el residu mutat en la posició en les seqüències alineades amb la d'estudi.

Taula2

Treball	N.Proteïne	N.Patològic	Origen	N. Neutre	Origen	Validació Creuada	Mètode predicció
Sunyaev et al.(Sunyaev et al., 2000b).	--	551	SWP	86	SWP & OMIM	no	Regles empíriques; Dades estructurals i evolutives
				250	HSSP 95%		
Wang & Moult(Wang & Moult, 2001).	23	262	dbSNP&OMIM	42	Halushka(Halushka, 1999). & Cargill(Cargill et al., 1999).	no	Regles empíriques; Basades en dades estructurals
Ng & Henikoff(Ng & Henikoff, 2001).	3	LAC 1750	Experiments de mutagènesi dirgida	LAC 2254	Experiments de mutagènesi dirgida	no	Optimització llindar de probabilitats d'aparició d'aa en columna d'alineament
		HIV 225		HIV 111			
		T4 638		T4 1377			
Chasman & Adams(Chasman & Adams, 2001).	2	LAC 1750	Experiments de mutagènesi dirgida	LAC 2254	Experiments de mutagènesi dirgida	si	Model Probabilístic; Determinació de llindar
		T4 638		T4 1377			
Saunders & Baker (Saunders & Baker, 2002).	3	LAC1166	Experiments de mutagènesi dirgida	LAC 2255	Experiments de mutagènesi dirgida	si	Llindar optimitzat sobre dades
		HIV 159		HIV 111			
		T4 175		T4 1340			
	--	191	SWP & OMIM	87	SWP & OMIM		

VIII. BIBLIOGRAFIA DEL CAPÍTOL

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31, 365-70.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q. & Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22, 231-8.

Chasman, D. & Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* 307, 683-706.

Collins, F. S., Brooks, L. D. & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* 8, 1229-31.

Collins, F. S., Guyer, M. S. & Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science* 278, 1580-1.

Cox, K., Watson, T., Soultanas, P. & Hirst, J. D. (2003). Molecular dynamics simulations of a helicase. *Proteins* 52, 254-62.

El-Bastawissy, E., Knaggs, M. H. & Gilbert, I. H. (2001). Molecular dynamics simulations of wild-type and point mutation human prion protein at normal and elevated temperature. *Journal of Molecular Graphics and Modelling* 20, 145-154.

Ellis, R. J. (2001). Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr Opin Struct Biol* 11, 114-9.

Ellis, R. J. & Minton, A. P. (2003). Cell biology: join the crowd. *Nature* 425, 27-8.

Fay, J. C. & Wu, C. I. (2001). The neutral theory in the genomic era. *Curr Opin Genet Dev* 11, 642-6.

Fay, J. C. & Wu, C. I. (2003). Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet* 4, 213-35.

Fay, J. C., Wyckoff, G. J. & Wu, C. I. (2001). Positive and negative selection on the human genome. *Genetics* 158, 1227-34.

Ferrer-Costa, C., Orozco, M. & de la Cruz, X. (2002). Characterization of Disease-associated Single Amino Acid Polymorphisms in Terms of Sequence and Structure Properties. *J Mol Biol* 315, 771-86.

Fersht, A. R. & Serrano, L. (1993). Principles of protein stability derived from protein engineering experiments. *Current opinion in structural biology* 3, 75-83.

Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins* 23, 566-579.

Futatsugi, N. & Tsuda, M. (2001). Molecular dynamics simulations of Gly-12-->Val mutant of p21(ras): dynamic inhibition mechanism. *Biophys J* 81, 3483-8.

Graur, D. & Li, W.-H. (2000). *Fundamentals of molecular evolution*. 2nd edit, Sinauer Associates, Sunderland, Mass.

Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M., Brody, L. C., Wang, D., Lander, E. S., Lipshutz, R., Fodor, S. P. & Collins, F. S. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet* 22, 164-7.

Halushka, M. K. (1999). Patterns of single-nucleotide polymorphisms in candidate genes regulating blood-pressure homeostasis. *Nat Genet* 22, 239-247.

Henikoff, S. & Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins* 17, 49-61.

Herrick, J. B. (1910). Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. *Arch. Intern. Med.* 6, 517-521.

Hua, Q. X., Shoelson, S. E., Inouye, K. & Weiss, M. A. (1993). Paradoxical structure and function in a mutant human insulin associated with diabetes mellitus. *Proc. Nat. Acad. Sci.* 90, 582-586.

Huisman, T. H. J., Carver, M. F. H. & Efremov, G. D. (1996). *A syllabus of human hemoglobin variants*, The sicle cell anemia foundation, Augusta.

Ingram, V. (2004). The sickle cell story ca. 1956. www.ergito.com.

Kimura, M. & Takahata, N. (1983). Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. *Proc Natl Acad Sci U S A* 80, 1048-52.

Lander, E. S. (1996). The new genomics: global views of biology. *Science* 274, 536-9.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Lercher, M. J. & Hurst, L. D. (2002). Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? *Gene* 300, 53-8.

Mao, C., Zhou, M. & Uckun, F. M. (2001). Crystal structure of Bruton's tyrosine kinase domain suggests a novel pathway for activation and provides insights into the molecular basis of X-linked agammaglobulinemia. *J. Biol. Chem.* 276, 41435-41443.

Martin, A. C., Facchiano, A. M., Cuff, A. L., Hernandez-Boussard, T., Olivier, M., Hainaut, P. & Thornton, J. M. (2002). Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. *Hum Mutat* 19, 149-64.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405, 442-51.

Matthews, B. W. (1987). Genetic and structural analysis of the protein stability problem. *Biochemistry* 26, 6885-8.

Matthews, B. W. (1993). Structural and genetic analysis of protein stability. *Annu Rev Biochem* 62, 139-60.

Matthews, B. W. (1995). Studies on protein stability with T4 lysozyme. *Adv Protein Chem* 46, 249-78.

McKusick-Nathans Institute for Genetic Medicine, J. H. U. B., MD) and National Center for Biotechnology Information, National Library of Medicine, (Bethesda, MD). (2000). Online Mendelian Inheritance in Man, OMIM (TM).

Minton, A. P. (2000). Implications of macromolecular crowding for protein assembly. *Curr Opin Struct Biol* 10, 34-9.

Mirkovic, N., Marti-Renom, M. A., Weber, B. L., Sali, A. & Monteiro, A. N. (2004). Structure-based assessment of missense mutations in human BRCA1: implications for breast and ovarian cancer predisposition. *Cancer Res* 64, 3790-7.

Ng, P. C. & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res* 11, 863-74.

Ng, P. C. & Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 12, 436-46.

Pauling L, I. H., et al. (1949). Sickle cell anemia a molecular disease. *Science* 110, 543-8.

Perryman, A. L., Lin, J. H. & McCammon, J. A. (2004). HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: possible contributions to drug resistance and a potential new target site for drugs. *Protein Sci* 13, 1108-23.

Poteete, A. R., Rennell, D. & Bouvier, S. E. (1992). Functional significance of conserved amino acid residues. *Proteins* 13, 38-40.

Poteete, A. R., Rennell, D., Bouvier, S. E. & Hardy, L. W. (1997). Alteration of T4 lysozyme structure by second-site reversion of deleterious mutations. *Protein Sci* 6, 2418-25.

Ramensky, V., Bork, P. & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 30, 3894-900.

Ramírez-Alvarado, M. & Regan, L. (2002). Does the location of a mutation determine the ability to form amyloid fibrils? *J Mol Biol* 323, 17-22.

Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol* 222, 67-88.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* 273, 1516-7.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E.

R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. & Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928-33.

Salisbury, B. A., Pungliya, M., Choi, J. Y., Jiang, R., Sun, X. J. & Stephens, J. C. (2003). SNP and haplotype variation in the human genome. *Mutat Res* 526, 53-61.

Santibanez-Koref, M. F., Gangeswaran, R., Santibanez-Koref, I. P., Shanahan, N. & Hancock, J. M. (2003). A phylogenetic approach to assessing the significance of missense mutations in disease genes. *Hum Mutat* 22, 51-8.

Saunders, C. T. & Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol* 322, 891-901.

Savitt, T. L. & Goldberg, M. F. (1989). Herrick's 1910 case report of sickle cell anemia: the rest of the story. *J.A.M.A.* 261, 266-271.

Shortle, D. (1992). Mutational studies of protein structures and their stabilities. *Q Rev Biophys* 25, 205-50.

Steward, R. E., MacArthur, M. W., Laskowski, R. A. & Thornton, J. M. (2003). Molecular basis of inherited diseases: a structural perspective. *Trends Genet* 19, 505-13.

Stockner, T., Sterk, H., Kaptein, R. & Bonvin, A. M. (2003). Molecular dynamics studies of a molecular switch in the glucocorticoid receptor. *J Mol Biol* 328, 325-34.

Sunyaev, S., Hanke, J., Brett, D., Aydin, A., Zastrow, I., Lathe, W., Bork, P. & Reich, J. (2000a). Individual variation in protein-coding sequences of human genome. *Adv Protein Chem* 54, 409-37.

Sunyaev, S., Kondrashov, F. A., Bork, P. & Ramensky, V. (2003). Impact of selection, mutation rate and genetic drift on human genetic variation. *Hum Mol Genet* 12, 3325-30.

Sunyaev, S., Ramensky, V. & Bork, P. (2000b). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* 16, 198-200.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A. S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum Mol Genet* 10, 591-7.

Sunyaev, S. R., Lathe, W. C., 3rd, Ramensky, V. E. & Bork, P. (2000c). SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet* 16, 335-7.

Syvanen, A. C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* 2, 930-42.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science* 291, 1304-51.

Vetrie, D., Vorechovsky, I., Sideras, P., Holland, J., Davies, A., Flinter, F., Hammarstrom, L., Kinnon, C., Levinsky, R., Bobrow, M., Smith, C. I. E. & Bentley, D. R. (1993). The gene involved in X-linked agammaglobulinaemia is a member of the src family of protein-tyrosine kinases. *Nature* 361, 226-233.

Vitkup, D., Sander, C. & Church, G. M. (2003). The amino-acid mutational spectrum of human genetic disease. *Genome Biol* 4, R72.

Wang, Z. & Moult, J. (2001). SNPs, protein structure, and disease. *Hum Mutat* 17, 263-70.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T.,

Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-62.

Wong, G. K., Yang, Z., Passey, D. A., Kibukawa, M., Paddock, M., Liu, C. R., Bolund, L. & Yu, J. (2003). A population threshold for functional polymorphisms. *Genome Res* 13, 1873-9.

Zhao, Z., Fu, Y. X., Hewett-Emmett, D. & Boerwinkle, E. (2003). Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene* 312, 207-13.

Zuckerkandl, E. & Pauling, L. (1962). Molecular disease, evolution and genetic heterogeneity. In *Horizons in Biochemisty* (B., M. & Pullman, B., eds.), pp. 189-225. Academic Press, London.

[*Aquesta pàgina ha estat deixada en blanc intencionadament*]